NBER WORKING PAPER SERIES

SHORT RUN IMPACTS OF ACCOUNTABILITY ON SCHOOL QUALITY

Jonah E. Rockoff
Lesley J. Turner

Working Paper 14564
http://www.nber.org/papers/w14564

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2008

We would like to thank Jesse Margolis, Raji Chakrabarti, David Figlio, Miguel Urquiola, and seminar participants at the New York City Department of Education, Cornell, Yale, Wharton, and the NBER Education Program meetings for their very helpful comments. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Short Run Impacts of Accountability on School Quality
Jonah E. Rockoff and Lesley J. Turner
NBER Working Paper No. 14564
December 2008
JEL No. H52,H75,I21,I28,L38

ABSTRACT

In November of 2007, the New York City Department of Education assigned elementary and middle schools a letter grade (A to F) under a new accountability system. Grades were based on numeric scores derived from student achievement and other school environmental factors such as attendance, and were linked to a system of rewards and consequences. We use the discontinuities in the assignment of grades to estimate the impact of accountability in the short run. Specifically, we examine student achievement in English Language Arts and mathematics (measured in January and March of 2008, respectively) using school level aggregate data. Although schools had only a few months to respond to the release of accountability grades, we find that receipt of a low grade significantly increased student achievement in both subjects, with larger effects in math. We find no evidence that these grades were related to the percentage of students tested, implying that accountability can cause real changes in school quality that increase student achievement over a short time horizon. We also find that parental evaluations of educational quality improved for schools receiving low accountability grades. However, changes in survey response rates hold open the possibility of selection bias in these complementary results.

Jonah E. Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

Lesley J. Turner
Department of Economics
Columbia University
1022 International Affairs Building
420 West 118th Street
New York, NY 10027
ljt2110@columbia.edu

Theoretical and empirical work by economists points out both the promise and pitfalls of school accountability systems. These systems aim to place pressure on schools to improve a set of quantifiable outcomes, often student achievement as measured by standardized tests. While several papers have documented positive benefits of accountability on student achievement, researchers have also found that the incentives and consequences of such systems may cause schools to take actions that improve accountability measures without raising actual student achievement (e.g., classifying low performing students into programs that exempt them from testing).[1] However, the literature on school accountability is still quite new, and there is much to learn about the various ways in which accountability systems affect educational outcomes and the behavior of students, teachers, and school administrators.

While the No Child Left Behind Act (NCLB) placed school accountability at the forefront of educational policy in the U.S., various states and cities have implemented their own systems of accountability, many of which preceded and go beyond the structure of NCLB to provide schools with incentives to increase student achievement. The New York City Department of Education (hereafter the DOE) launched its own system in the fall of 2007. Like other accountability systems, the DOE evaluated schools according to a series of continuous metrics, but each school was assigned a letter grade from A to F based on sharp cutoffs in these metrics. The system offered increases in per pupil funding and bonuses for principals in schools that did well. For schools that performed poorly, the system threatened leadership change and possible closure, required a series of formal corrective actions, and allowed students to transfer to other schools. Importantly, these rewards and consequences were linked to schools' grades,

[1] See, for example, Ladd and Zelli (2002), Jacob and Levitt (2003), Figlio and Winicki (2005), Hanushek and Raymond (2005), Jacob (2005), Chakrabarti (2006, 2008), Cullen and Reback (2006), Figlio (2006), Figlio and Getzler (2006), Figlio and Rouse (2006), Chiang (2007), Rouse et al. (2007), Neal and Schanzenbach (2007), Kreig (2008), and Reback (forthcoming).

rather than the underlying continuous metrics that determine them. The discontinuities inherent in the assignment of grades present an opportunity to study the short run effects of accountability on student achievement in a way that can uncover the causal impacts of accountability grades on student and school outcomes.

We use aggregate, school level data on student achievement to determine whether grade assignments in the fall of 2007 had an impact on student achievement in the early months of 2008. Specifically, grades were released to schools in late September and made public in early November, while math tests were administered in March and English Language Arts (hereafter English) tests were given in January. Thus, school administrators had between four and six months to respond to their grade assignment. We find positive, statistically significant, and economically meaningful impacts on student achievement in math for schools that received a grade of F or D and on English achievement for school that received an F. We also find evidence that the impact of accountability grades on English achievement was larger for schools whose prior average scores fell within the bottom half of all schools in the city. Importantly, we find no relationship between accountability grades and students' probability of being tested, one measure of possible gaming. Last, but not least, we examine a set of outcomes from annual surveys of parents, teachers, and students. We find complementary evidence that evaluations of school quality, particularly by parents, rose significantly for schools receiving low accountability grades. While not conclusive, these results suggest the impacts we document are driven by genuine improvement among schools receiving low grades.

The paper proceeds as follows. We describe the DOE accountability system in Section 2. Section 3 describes the data and provides descriptive statistics. Section 4 lays out our empirical strategy and discusses results from graphical and regression analyses. Section 5 concludes.

2. School Progress Reports in New York City

In 2006, the New York City DOE implemented a new accountability system based on annual data encompassing various aspects of school performance. In this paper, we focus on the central piece of the system—school progress reports. The main feature of the progress report was a letter grade, ranging from A to F, which was based on several continuous measures of success. The report also provided a summary of the factors contributing to the school's grade, a "quality review" score based on a qualitative evaluation, in which the school was ranked as "Well Developed," "Proficient," or "Undeveloped," and the school's NCLB status. A school's accountability grade, quality review, and NCLB status were independently determined. Progress reports were generated for elementary, middle, and high schools throughout the city, but due to data limitations, we exclude high schools from our analysis.

Accountability grades were based on student achievement test scores, attendance, and evaluations of the school environment from annual surveys of students, parents, and teachers. Quality review scores were determined by an on-site review by an "experienced educator" and were based on "the quality of efforts taking place at the school to track the capacities and needs of each student, to plan and set rigorous goals for each student's improved learning, to focus the school's academic practices and leadership development around the achievement of those goals, and to evaluate the effectiveness of [these] plans and practices…" (Educator Guide, New York City Progress Report). While the same test scores used in setting accountability grades determined a school's NCLB status, a different set of calculations were employed. In particular, a school's NCLB status was based on the fraction of students in the school (and in various

subgroups) scoring above a threshold.[2] Accountability grades, as we explain, rely on a broader set of information.

Accountability grades were based on an overall score determined by a school's performance in the three main report elements – school environment (15 percent of the overall score), student performance (30 percent), and student progress (55 percent).[3] Schools also received additional credit for sizeable achievement gains ("exemplary student progress") within particular student subgroups (e.g., English Language Learners). A number of components contribute to the score a school received for each report element.[4] A school's environment score was determined by the responses from surveys administered to students (in grades 6 and above), their parents, and the teachers at the school, as well as student attendance rates.[5] The performance score was determined by several measures of student achievement, as measured by the state math and English examinations, while the progress score was based on changes in individual student achievement on the these examinations. (In New York State, students in grades 3 to 8 are tested annually in math and English as part of compliance with NCLB.) A school's overall score equals the weighted average of the environment, progress, and performance scores and any additional credit for exemplary student progress.

---

[2] We do not focus on NCLB here, although there are many similar issues between it and the DOE system. For an overview of NCLB related topics, see Peterson and West (2003).

[3] The DOE provides extensive information on the formulation of school accountability grades. See schools.nyc.gov/NR/rdonlyres/DF48B29F-4672-4D16-BEEA-0C7E8FC5CBD5/27499/EducatorGuide_EMS.pdf

[4] There are four to six components for each of the three categories. The components of the environment score are the school's attendance rate and indices of school safety, academic quality, student engagement, and communication taken from an annual survey of parents, teachers, and students. The components for student performance are the percentage of students that were proficient (i.e., level 3 or 4) on the state English (math) examinations and the median score received by students on the English (math) examinations. The components of student progress are the percentage change in individual students' English (math) scores, the average change in proficiency among all students in the school, and the average change in the proficiency of the lowest third of students, as determined by students' previous year proficiency ratings. In the case of mid-year student transfers, the credit that the sending and receiving schools receive for a student's performance is determined by the portion of time a student spent at each school during the period between the current and previous state examinations.

[5] See http://schools.nyc.gov/Accountability/SchoolReports/Surveys/2007survey.htm for more information.

Schools earned additional credit for individual achievement gains made within five subgroups of students: students with performance in the lowest third of all students citywide who were Hispanic, Black, or other ethnicities, and students in English Language Learner (ELL) or Special Education programs. To receive additional credit, the percentage of students within a subgroup whose achievement gains exceeded a set magnitude must fall within the top 40 percent of all schools of its type (e.g., elementary, middle, or K-8 school) citywide.[6] Overlap across student groups is allowed. For example, if an ELL student was in the lowest third citywide, he/she would be counted in the calculation of additional credit for both groups. Additional credit was only given if there were at least 20 students in the subgroup; if fewer than 20, Hispanic or Black students would be aggregated with students of other ethnicities. Schools received 0.75 additional points for having gains within a particular group that fell within the top 40 percent of schools of its type and an additional 0.75 points (for a total of 1.5) if the gains were within the top 20 percent.

A school's score for each report element (environmental, performance, and progress) was determined both by that school's performance relative to all schools in the city of the same type and relative to a group of schools with similar students.[7] Each school was assigned a "peer index" based on either the composition of the student body (elementary and K-8 schools) or the performance of current students on state exams taken prior to their arrival at the school (middle

[6] Specifically, the DOE measures the fraction of students whose scores increase by "half of a performance level or more." Performance levels, which range from 1 to 4.5, are simply a rescaled version of the scaled score that is more familiar to DOE personnel.

[7] Schools are placed in one of these three categories (elementary, middle, or K-8 school) depending on their grade level structures. Middle school structures are grades 5–8, 6–8, and 6–12 (excluding 9-12 graders), K-8 school structures are K–7, K–8, and K–12 (excluding 9-12 graders), and elementary school structures are all other combinations serving grades lower than 7. A small number of schools (10 of 581 elementary and 1 of 116 K-8) will be classified differently for the 2007-2008 school progress reports due to changes in grade offerings. We use 2006-2007 classifications throughout our analysis.

schools).[8] Within each school type, schools were ordered according to their peer index and compared with the 20 schools just above and the 20 schools just below.[9] Thus, each school was designated a unique peer group.

For each report element, schools received a "city horizon" and a "peer horizon" score, based on its relative standing within the comparison group.[10] The weighted average of the school's relative performance in each component—with performance relative to peer schools given double the weight of relative performance citywide—determined a school's score for the particular element. The weighted sum of the scores of each of the three elements (but prior to receiving additional credit for exemplary student progress) was then calculated for each school. These pre-additional credit scores could range from 0 to 100. Additional credit is then added to produce the school's overall score.

Within each school type, the DOE ranked schools by their pre-additional credit scores and assigned each school a percentile. These percentiles were then used to determine the cutoff scores between accountability grades for each type of school. The cutoff score to receive an A was set at the 85th percentile, B at the 45th percentile, C schools at the 15th percentile, and D at the 5th percentile. However, schools were assigned grades based on whether their overall score,

[8] Elementary/K-8 schools received a peer index score ranging from 0 to 100 determined by the percentage of students eligible for free lunch (40 percent of the score), the percentage of students that are Black or Hispanic (40 percent), and the percent of the school categorized as Special Education students (10 percent) or English Language Learners (10 percent). Middle schools were assigned a peer index score ranging from 1 to 4.5 based on the average performance level received by currently enrolled students on their fourth grade state exams.

[9] Schools at either end of the distribution of the peer index scores were assigned a group of less than 40 schools – among the 985 schools we examine, 62 percent had a full group of 40 peer schools. All schools had at least 20 peer schools. In some cases, peer schools included charter schools. Charter schools that were at least two years old and had test score results for third and fourth graders received a progress report. However, accountability grades received by charter schools are not comparable with those received by other schools, as the environment category score was only based on attendance.

[10] The horizon scores were based on the ratio of a school's score minus the city (or peer) "minimum score" to the city (peer) "maximum score" minus the city (peer) minimum. The "maximum" ("minimum") score is actually defined as the mean plus (minus) two standard deviations, and is thus closer to a z-score than a percentile ranking. Progress reports provide some language that explains this methodology and provide a sample calculation to help readers further understand it.

including additional credit, exceeded these thresholds. Of the 985 schools we examine, approximately 75 percent received some additional credit. The impact of additional credit points was not negligible – 161 schools received a higher grade due to additional credit. Of these schools, 6 moved from an F to a D, 22 moved from a D to a C, 57 moved from a C to a B, and 76 moved from a B to an A.

Thus, the percentage of schools receiving each accountability grade does not precisely correspond to the original percentile cutoffs. Specifically, among the 985 schools we examine, 23 percent received A's, 38 percent received B's, 26 percent received C's, 9 percent received D's, and 4 percent received F's. Figure 1 shows the relationship between accountability grades and overall scores. There are clear discontinuities in the assignment of grades as we move up the continuous distribution of overall scores.

Figure 2 displays a timeline of events that occurred as the accountability system was developed and implemented. Progress reports were first provided to principals on September 24th, 2007, and were released to the general public on November 5th. However, principals were first informed of the progress report methodology in April of 2007. At this time, principals also received a pilot progress report with numeric scores based on achievement data from 2005 and 2006. Nevertheless, these pilot reports did not contain a letter grade, and principals were not told how scores would translate into grades. These reports also lacked other important pieces of information that would affect their actual progress reports, such as environmental scores and peer groups. Thus, we believe it is highly unlikely that schools could have predicted the grades they received in the fall of 2007 with the limited amount of information they were given earlier that spring. Anecdotally, some principals receiving low grades were surprised (New York Times, November 4, 2007). School closures that were based on poor performance were announced in

December of 2007. Tests in English and math were taken on, respectively, January 8-17 and March 4-11. The environmental survey of parents, students, and teachers was administered between March 12 and April 18, well before any test score results were released.

There are several reasons why accountability grades may create pressure for schools to raise student achievement. First, the system generated consequences for schools performing poorly. Schools receiving an F grade and a poor quality review rating are likely to undergo a leadership change or even face closure. Indeed, 7 schools receiving an F and 2 schools receiving a D were told in December of 2007 that they would be closed immediately or phased out after the school year 2007-2008.[11] Thus, soon after the receipt of accountability grades, it was made clear to schools that the threat of closure for poor performance was real. Additionally, 17 percent of the *remaining* F school principals (and 12 percent of the remaining D school principals) did not return in the school year 2008-2009, relative to 9 percent of principals in schools receiving a C, B, or A grade.[12] Students in F schools were also eligible to transfer through a special application process which occurred in the summer of 2008, and all schools receiving D or F grades were required to implement formal "school improvement measures and target setting" and would be subject to leadership change if they continued to receive low grades.[13] Schools receiving a C grade for three years would also face these consequences. While some accountability systems (e.g., Florida's system) offer additional funding for school

[11] One of the schools we consider to have closed was serving grades 6-12 and now only serves grades 9-12. Our results are robust to the exclusion of the schools that knew they were to be closed in December. Although F and D schools were at risk for school closure or principal removal, these decisions were determined on a case by case basis by the Chancellor, and were not based on an explicit formula. Principals and teachers in schools that close do not face unemployment. They can search for another position within the district through normal channels, and, if not successful, teachers work as substitutes throughout the city and principals are assigned to serve as additional administrators in schools or central district offices.

[12] We do not know for certain which principals were removed due to progress report grades, but the higher rate of departure of principals in D and F schools is consistent with the provisions of the accountability system.

[13] See http://schools.nyc.gov/Accountability/SchoolReports/ProgressReports/Consequences/default.htm.

improvement to poor performing schools, F and D schools in New York did not receive any additional funds.

Second, the system linked financial rewards to school performance. Schools that received an A grade and a "Well Developed" quality review rating received a funding increase for the following school year of roughly $33 per student, which can be used at the school administrator's discretion.[14] These payments totaled $3.4 million in the school year 2007-2008. Schools that received an A or B grade and a "Well Developed" or "Proficient" quality review rating were also eligible for payments of $1,500 to $3,000 per student per year for any student accepted as a transfer from a school that received an F or a school not in good standing under NCLB. Last, but not least, principals of schools with an overall score among the top 20 percent citywide (within each type of school) and a "Well Developed" or "Proficient" rating for their quality review are eligible to receive monetary bonuses of $7,000 to $25,000.[15]

The publicity surrounding the accountability grades may also generate pressure (e.g., from parents) for schools receiving low grades to improve their performance. News reports at the time (see New York Times, November 4, 6, and 7, 2007) provide clear indication that the release of progress report grades captured the attention of principals and parents alike, although reactions were mixed among both high and low rated schools. Principals and parents worried that the progress reports put too much emphasis on testing and did not accurately reflect their school's quality, but also emphasized the incentives schools to "keep up" with peer schools. Although many found the methodology involved with assigning grades complicated, the status of

[14] Expenditure per pupil in the DOE for the school year 2005-2006 (the last year data is available on the New York State department of education website) for general education students was $9,526 (see http://www.emsc.nysed.gov/irts/reportcard/2007/supplement/300000010000.pdf). Assuming 5 percent growth in spending, the $33 bonus would amount to a 0.3 percent budget increase.

[15] The progress reports released in November, 2007 did not result in bonuses; they will be given out for the first time in the fall of 2008 and will depend on progress reports based on 2007-2008 performance. The top 1 percent of all principals receive $25,000, the next 4 percent receive $17,000, the next 5 percent receive $12,000 and the next 10 percent receive $7,000. Assistant principals get half of the bonus that their principals receive.

receiving a high grade and the consequences attached to receiving a failing grade – as severe as closure or the removal of the current principal – were also very clear.

3. Data

Our primary source of data is a set of files publicly available on the DOE website. The first two files provide achievement test results from 2006 to 2008 at the school-grade cell level for every school in the DOE serving grades 3 to 8. Students in these grades are tested annually in English and math in accordance with NCLB. These data include the number of students tested and the average "scale score," by year and grade level; the average scale score, under the assumptions of item response theory, is a valid measure of group average achievement. The third file from the DOE contains the accountability grade assigned to each school, the overall score used to assign that grade, the elements of the overall score, and the school's NCLB status.

There are 1092 elementary, K-8, and middle schools with 2008 student math and English achievement data. We exclude the 40 schools belonging to District 75 (which primarily serves special education children) and the 25 schools that did not have math or English achievement data for the school year 2006-2007.[16] Of the remaining schools, an additional 42 did not receive an accountability grade for various reasons. For example, one of these schools specializes in serving recent immigrants for one year, making it impossible to measure changes over time in achievement for their students. A number of other schools not receiving grades were already in the process of closing. Our final sample consists of 985 schools and represents 90 percent of the schools with 2007-2008 achievement data.

We present summary statistics by accountability grade in Table 1. The distribution of grades is similar across elementary, K-8, and middle schools. Enrollment, however, is

[16] Of these schools, 24 were not assigned accountability grades and the one school that did receive a grade did not have 2007 math achievement data.

noticeably lower in schools receiving an F, D, or A grade, and the fraction enrolled in tested grades (3-8) is also particularly low in F schools.[17] We find that schools receiving an A are more likely to be in good standing under the NCLB accountability system than schools receiving a B or lower, yet there are no other noticeable differences in NCLB status by accountability grade among schools not receiving an A.

In order to further characterize schools with different accountability grades, we merged the DOE data with student level data from the school year 2006-2007 covering all students in grades 3 to 8. Student characteristics bear a noticeable relationship with accountability grades (Table 1). Higher grades are associated with fewer students receiving free lunch, fewer special education students, fewer black students, and more white and Asian students. Interestingly, we see weaker relationships between accountability grades and the fraction of Hispanic students and English Language Learner students.

The middle of Table 1 presents average student achievement outcomes by accountability grade for the school years 2006-2007 and 2007-2008. As we might expect, 2006-2007 achievement outcomes increases monotonically with progress report grades. The gap between the test score averages for A and F schools in 2006-2007 is 17.9 points in English and 23.8 points in math. In the school year 2007-2008, the monotonic relationship between accountability grades and test scores still is present. However, while average test scores improved for schools receiving every grade, the greatest improvements were made by schools receiving lower grades. The gap between the test score averages for A and F schools shrank to 12.8 points in English and 19.1 points in math. While this does not necessarily imply that the distribution of achievement

[17] There are two plausible reasons for this pattern. One is the fact that variance in test score outcomes will be greater for smaller populations of tested students (see Kane and Staiger (2002)), making them more likely to end up with either very high or low measured performance. The second is that school size and/or grade composition are related to other characteristics that are indicative of high or low performance. Distinguishing these explanations is beyond the scope of this paper.

among DOE schools decreased between the 2006-2007 and 2007-2008 school years, we find that the standard deviation of achievement across the schools in our sample fell from 17.2 to 15.3 points in English and from 21.2 to 19.5 in math. To give a better sense of this compression, we plot kernel densities of school average scaled scores by year (Figure 3). This graph shows that test scores among schools in New York improved at nearly every percentile in both subjects, but noticeably greater gains were made below the top quartile.

The bottom half of Table 1 shows the continuous metrics underlying the accountability grade and quality review ratings. For ease of exposition, we normalize the peer indices within school type to have a mean of zero and standard deviation of one, and reverse the sign of the elementary and K-8 school peer indices (which are based on percentage of students by ethnicity and program participation) so that they are positively correlated with school average achievement levels. Not surprisingly, the average overall score and scores for the report elements increase monotonically as we move from F to A. Schools receiving lower grades also had lower peer indices, indicating that these schools served more disadvantaged students (for elementary/K-8 schools) or students who had scored poorly on the achievement tests in the past (for middle schools). Quality review ratings are also correlated with the accountability grades, though these measures have no mechanical dependence. Among F schools, 17 percent were rated as "Undeveloped" and 17 percent were rated "Well Developed," compared with 2 percent rated "Undeveloped" and 48 percent rated "Well Developed" among A schools.

To serve as a point of comparison, we also examine school characteristics according to NCLB status to determine whether similar trends exist across schools in good standing, those found in need of improvement, and schools planning or undergoing restructuring for poor performance (see Table 2). As mentioned above, NLCB status is based on the same achievement

tests but uses a very different formula, looking only at the percentage of students scoring above a passing threshold. We find that demographic differences between schools in good standing and those planning or currently in the process of restructuring are as large, if not greater than those found when we examine schools according to accountability grade. Given NCLB performance is based only on current performance, this finding is not surprising. School demographic characteristics and average test scores (which may both capture unobservable characteristics, such as resources at home and within the school) are likely to have a higher correlation than demographics and individual student progress (which may have a stronger correlation with processes within a school). One exception is that the portion of the student body that is black varies little across NCLB status while the portion that is Hispanic or English Language Learner varies considerably.

4. Empirical Methods and Results

The empirical methods we employ are very much in the spirit of previous work on the impacts of school accountability grades (e.g., Figlio and Lucas (2004), Rouse et al. (2007), Mizala and Urquiola (2008)) and other work using regression discontinuities to identify the impact of educational policies (e.g., van der Klauww (2002), Jacob and Lefgren (2004), Chay et al. (2005)). Specifically, we use the discontinuous relationship between accountability grades and the numeric inputs that determine the grades to compare the subsequent outcomes in schools that received different accountability grades but were otherwise similar. To estimate this impact, we use a reduced form regression specification represented by Equation 1.

$$(1)\quad A_{jt} = \alpha + \lambda_G D_{jt}^G + \beta f(P_{jt}) + \varepsilon_{jt}$$

Here, A_{jt} is the average achievement of students in school j and year t, D_{jt} is an indicator for the accountability grade (G) assigned to the school, P_{jt} are the continuous measures used to

determine the accountability grade (i.e., environmental, performance, and progress scores, additional credit, and peer index), and ε_{jt} is an idiosyncratic noise term. We include a quartic in P_{jt}; including higher order polynomials does not noticeably change our results. Also, because the cutoffs for accountability grades and the scaling of the peer index differed across the three school types, we include indicators for school type and interactions of school type with the quartic in the continuous measures P_{jt} in all of our specifications. Our identification strategy is based on the notion that accountability grades are exogenous and uncorrelated with the error term ε_{jt} once we have conditioned on all the factors used to determine the accountability grade. Under this assumption, estimates of the parameters λ_G will give us unbiased measures of the impact of school accountability grades on student achievement.[18]

As with all studies that use the regression discontinuity approach, our method identifies the impact of differences in the behaviors of schools with overall scores that place them on either side of the margin between two accountability grades. For example, we compare subsequent outcomes of schools who receive a "low C" with schools receiving a "high D." If both types of schools feel the same pressure to improve performance, then we might find little difference in their subsequent outcomes. However, such a finding does not mean that the accountability system does not create pressure to improve performance or lead to improved student outcomes. Rather, it means there is no discontinuity in this effect between schools with similar overall scores but different grades. Thus, the RD methodology is a limited test of the accountability system.

[18] Since test scores are inherently noisy measures of ability, one might worry that estimated impacts of receiving a poor grade might be biased by regression to the mean (e.g., schools receiving very high scores had an unexpectedly good shock to test scores while schools receiving very low scores had a bad shock). However, the regression discontinuity methodology will avoid this concern as long as regression to the mean can be specified as a flexible continuous function of the variables that determine school grades (Chay et al. (2005)).

One important issue with regard to research on accountability is student mobility. It may be the case that any differences in test scores and gains that appear to be due to the accountability grades are actually driven by higher achieving students transferring to a school with a better grade. With the aggregated data currently used in this paper, we cannot test this directly. However, the timing of the grade announcements and subsequent student testing greatly reduces concerns regarding student mobility. The accountability grades were released in November of 2007 and the tests were taken within the same school year. Special transfers out of F schools did not occur until the summer of 2008. Thus, any student transfers in response to accountability grades would need to be self-initiated and occur in the middle of the year. Such moves are likely to be viewed by parents as highly disruptive (see Hanushek et al. (2004)).

Additionally, we can examine mid-year transfer behavior using student level data covering the school years 2002-2003 through 2006-2007, which contain information on students' locations in October and June of each school year. We find that about 2.5 percent of students present in October were in a different school within New York City in June. Thus, there is limited scope for migration to impact our findings. Third, it is not unreasonable to think that the students most motivated to move to a better rated school would also be higher achieving than their classmates, thus, even if students transferred mid-year in response to a school's grade, any impact of migration is likely to bias us against finding positive effects of lower accountability grades on student achievement.

The literature on school accountability also raises the concern of gaming (e.g., Figlio and Winicki (2005), Cullen and Reback (2006), Figlio (2006), and Figlio and Getzler (2006)). While we cannot address all potential concerns regarding this issue, with aggregate data, we can test

whether accountability grades are associated with differences in the proportion of students taking math and English tests. We do not find any evidence that this is the case (see Section 4.4).

4.1 Graphical Analysis

Before proceeding to our regression analysis, we present a graphical depiction of our estimation strategy in Figures 4 and 5. First, we plot school average math and English scaled scores against the overall accountability score received by each school, using different symbols to distinguish schools that received different accountability grades. Then, we plot the residuals from regressions of scaled scores in math and English on inputs that determined the accountability grade (i.e. peer index, report element scores, and additional credit). Specifically, we allow for a quartic polynomial in each input and allow for different relationships within each type of school (i.e., elementary, K-8, middle). To aid with interpretation, each graph includes a line tracing the results of a locally weighted "Fan" regression (Fan and Gijbels (1997)) that provides a weighted average of performance at various levels of schools' overall scores, calculated separately within each group of schools that received the same grade. Breaks in the locally weighted regression line at the margins between accountability grades indicate a change in the performance of schools with similar overall scores but different grade assignments.

Figure 4 presents these graphs for 2006-2007 scaled scores. As student performance played a significant role in the calculation of grades, contributing 30 percent to the overall score, it would not be surprising if the overall score and the raw scaled scores were related. However, we see some interesting and unexpected patterns. Scaled scores in both subjects rise on average between each of the five grades, but within grades, scaled scores are increasing in overall score only for schools receiving grades of F or D. For schools receiving C and A grades, the relationship is fairly flat, and for schools receiving B grades there appears to be a slightly

negative relationship between scaled scores and the overall accountability score. We also see what appear to be significant breaks at the grade margins at every margin, which is unexpected. It is not clear to us why these breaks would occur, though it may simply be an artifact of an interaction between the manner in which grades were assigned and the cutoff values between grades, the fact that there is a large amount of variance in average test scores among schools receiving very similar overall accountability scores, and the relatively thin density of schools, especially among those receiving lower grades. Given the manner in which the cutoffs were determined (i.e., based on percentiles) and the fact that no school was assigned a grade for which their overall score did not warrant, we still regard the grade assignments as exogenous conditional on the inputs into the overall score.

The bottom panel of Figure 4 supports this notion. When we plot residuals from regressions that control for the overall score inputs, we find essentially no differences between schools receiving different grades, no noticeable trends within these groups of schools, and (consequentially) no major breaks at the margins between grades. Thus, when we control for the inputs used in assigning grades, the actual letter grades received by schools have no predictive power for 2006-2007 test results. The only detectable change at any margin is a small difference between F and D schools, where average scores for D schools are slightly higher.

Figure 5 displays the same information but using 2007-2008 scaled scores. The graphs of raw scores in the top panel show the same noticeable differences in average test scores, trends (both positive and negative), and breaks at the margin that were seen in the prior year. However, the bottom panel, which plots the residuals, looks quite different. For math scores, we can see noticeably higher test scores for F and D schools, and breaks at the F-D and D-C grade margins, but no differences or breaks at the margins for C, B, and A schools. For English scores, we see

higher scores among F schools and a break at the F-D margin, with no differences or breaks at the higher grades. These results indicate a positive impact on both English and math scores for schools on the margin of receiving an F and a D, and for math results, a positive impact for schools on the margin of receiving a D and a C. These graphs at the bottom panel of Figure 5 represent our essential findings. In the next section we present evidence from regression analysis that provides point estimates and standard errors on the qualitative findings from these graphs.

4.2. Regression Estimates of Impacts on Average Test Scores

In this section, we present results of regression specifications in the form of Equation 1. School average scaled scores are regressed on indicators for accountability grade and the inputs that determined the overall score. In Table 3, we first present results that examine test scores from the school year 2006-2007. We expect to find no significant differences in scaled scores across grades conditional on the inputs for the overall score. This is confirmed by the data; none of the indicator variables for grade are statistically significant, and tests for the equality of the coefficients between adjacent grades cannot be rejected (Table 3, Columns 1 and 2).

We see very different results when we examine test scores from 2007-2008. As foreshadowed by our graphical analysis, we find significantly higher test scores for F and D schools in math and F schools in English, conditional on our flexible controls for overall score inputs. A test of equality between the D and F coefficients can be rejected at the 3 percent level for math and the 8 percent level for English (Columns 3 and 6). The remaining columns in Table 3 provide two additional specifications. The first includes a quartic polynomial of the school's prior average scaled score as additional control variables. While this steps outside the set of variables that directly enter the accountability grade calculation, it further controls for any pre-existing differences between schools receiving different grades. One might be concerned in this

regard given that in the 2006-2007 test score regressions we find positive, though statistically insignificant, coefficients for F and D schools. Although adding these controls (Columns 4 and 7) causes a small reduction in the point estimates, it does not affect the significance of our initial findings that schools receiving F and D grades experienced an improvement in test scores. In fact, the addition of these controls reduces the standard errors considerably, and the negative point estimate for schools that received an A is now marginally significantly different than schools that received a B for both math and English performance (at the 7 and 9 percent level, respectively), suggesting that schools assigned a grade of B improved their scores relative to A schools.

The final specification drops a small number of schools that received an overall score either well below or well above the rest of the schools. Specifically, we drop 10 schools with an overall score below 15 or above 90 (more than two standard deviations from the average overall score; these include 5 A and 5 F schools). This has little impact on the results. Taking the point estimates from this final specification, we estimate that receipt of an F grade increased math and English scores by 2.1 and 1.8 scaled score points, respectively, relative to a receiving a D, and that receipt of a D grade increased math scores by 2.1 scaled score points (relative to a C). We also find some suggestive evidence that receipt of a B may have increased math and English scores by 1.3 and 0.8 points, respectively, relative to schools receiving an A.

One final issue is that our analysis includes the nine schools that learned in December of 2007 that they were to be closed or phased out after the end of the school year 2007-2008. One might think that these schools do not face any threat of further consequences and should not respond similarly do other schools, and one would be concerned if our results hinged on test score changes in these schools. We therefore repeat our analysis dropping these schools from

our sample. We generally find slightly larger point estimates for the impact of accountability on student achievement in F and D schools (results available upon request), suggesting that, if anything, the response among F and D schools was indeed greater among those not facing closure.

There are several ways we can gauge the magnitudes of these effects. First, we can compare the effects we find to the citywide standard deviation of changes in school average scale scores from 2007 to 2008, which were 5.7 points in math and 4.8 points in English. Thus, the impact of receiving an F, relative to a D, increased scores in math and English by nearly 0.4 standard deviations on this distribution, with a similar impact of receiving a D, relative to a C, on math scores. Second, note that the difference in average scale scores between C schools and F schools was approximately 11.8 points in math and 9 points in English. The gaps between A and D schools are roughly the same; 13.2 points in math and 9.6 points in English. Thus, our estimates suggest that the short run impact on achievement of students in schools receiving F grades was about 18 percent and 20 percent of the C-F gap in math and English, respectively. Additionally, the impact on D schools was about 16 percent of the A-D gap in math.

Finally, we can also judge these effects as the fraction of a student level standard deviation. While 2008 test score data at the student level are not yet available, we know that the standard deviations of math and English in 2007 were roughly 38 and 42 scale score points, respectively. Thus, the increases in math test scores for F and D schools (relative to C schools) were 0.1 and 0.05 standard deviations, and the increase in English test scores for F schools (relative to D schools) was 0.05 standard deviations. While these magnitudes are somewhat smaller than the estimated impact of receiving an F grade in the state of Florida on student performance one year later (Rouse et al. (2007)), they are economically significant and of similar

magnitude to other estimates of how improvements in school quality affect student achievement, such as attending a school with higher achievement levels (Hoxby and Weingarth-Salyer (2005), Hastings et al. (2007), and Hastings and Weinstein (2008)) or being assigned a highly experienced teacher (Kane et al. (2008).[19] Nevertheless, it is also important to note the possibility that the accountability system induced test score increases at low performing schools through the teaching of test-taking skills, rather than a true increase in the quality of math or English instruction.[20]

4.3 Heterogeneity in the Effects of Accountability Grades

In the results above, we estimated the average impact of accountability grades across all schools in the DOE. Here, we examine whether these effects vary across groups of schools for which we see a potential for different reactions to the accountability system, though, in each case, we have no firm hypothesis based on theory. First, we separate middle schools from schools serving lower grades (K-8 or elementary schools). While we have no prior as to which type of school would face greater pressure under the accountability system, we would note that, unlike middle schools, a large fraction of students in elementary/K-8 schools do not take standardized tests. On one hand, this may make it easier for elementary/K-8 schools to improve test scores, since they can focus on just a subset of their students, or it may be the case that it is easier to improve achievement in lower grades over a shorter time frame. On the other hand,

[19] Hoxby and Weingarth-Salyer (2005), Hastings et al. (2007), and Hastings and Weinstein (2008) estimate that moving elementary or middle school students to a school whose average test scores are a student-level standard deviation higher is expected to raise their test scores by about 0.15 to 0.5 standard deviations. To put this result into context, the gap in average achievement between F and C schools in New York is about 0.25 student level standard deviations in math, implying that moving students from F schools to C schools would raise their achievement by 0.04 to 0.12 standard deviations in expectation. Using data from New York City, Kane et al. (2008) find that elementary and middle school students assigned a highly experienced teacher (as compared to a rookie) are expected to have 0.08 student level standard deviations higher math achievement.

[20] A randomized evaluation of a short-run preparation program for the Scholastic Aptitude Test (Alderman and Powers (1980)) indicates an effect on SAT Verbal scores of roughly 0.08 standard deviations.

having non-tested grades may make schools less likely to place resources into teaching the tested material, since it only applies to a subset of their students. We then divide schools according to NCLB status, separating those that were in good standing under NCLB and those that were not. One might hypothesize that schools already under pressure from NCLB that receive a low grade may face greater pressure to improve or already be in the process of implementing measures to address short-comings in the face of NCLB-imposed consequences. On the other hand, the pressure already on these schools may mean that the DOE accountability system has little additional effect. Finally, we divided schools by whether their average scaled scores in the prior school year were higher than the median for their school type. The level of test scores in a school is often used as a proxy for school quality, and there is evidence that this measure is valued by parents (e.g., Black (1999)). Schools with high test score levels who received a low rating may have felt less pressure due to their already being thought of as effective. On the other hand, a low accountability grade for these schools may have served as a greater shock, and, in turn, may have caused a greater response.

To examine heterogeneity in the effect of accountability grades, we include in our regression interactions of accountability grade indicators with two indicator variables that separate schools into two groups (e.g., an indicator for being in good standing under NCLB and an indicator for not being in good standing), and drop the main effects of accountability grades. Thus, the coefficients we report represent the effect of receiving a particular grade for schools in one category and can be directly compared. We present results on heterogeneity in Table 4. For simplicity, we only show the results for 2007-2008 from specification that controls for a quartic in prior scale score in addition to the inputs into the overall score, but other specifications (i.e. dropping these controls, dropping outliers) resulted in similar findings.

We find little evidence of heterogeneity along these dimensions. The magnitude of the coefficients for receiving an F or D on math scores are somewhat larger for middle schools than for schools serving lower grades, however these differences are not statistically significant (Columns 1 and 2). For English scores, the coefficients are similar for both types of schools. We also do not see significant differences in the impacts on math or English scores when comparing schools by NCLB status (Columns 3 and 4). We do find some evidence of heterogeneity of effects on English scores for schools whose prior test scores were below or above the median. In particular, it appears that the estimated impact of receiving an F grade on English scores may be driven by improvements among schools with below median prior average scores. We can reject equality of effects at only the 22 percent level (perhaps because only 10 of the 42 F schools had prior average English scores above the median), but the point estimates for schools below and above median are, respectively, 2.45 and -0.02 scaled score points. Interestingly, we also find evidence that accountability grades above the F level had greater impacts on English scores for schools with prior scores above the median. Specifically, the point estimate for B schools (relative to C schools) is statistically significant at below the 5 percent level. These estimates suggest that, for these schools, test scores improved by about 1.3 scaled score points more for schools receiving a C grade (relative to those receiving a B). These results provide a potentially interesting story behind how accountability grades impacted student achievement in English. It may have been that case that, in order to impact student achievement, schools with low average achievement only reacted to receiving an F, while getting a C was sufficient motivation for schools with high average achievement. However, it is important to note that we do not find noticeable differences across these types of schools in the impact of

accountability grades on math test scores, so this interpretation should be taken with a fair amount of caution and this issue should be pursued further in future research.

4.4 Did Accountability Affect Testing Probabilities?

Without student level data, it is difficult to test many of the ways in which schools may try to improve their measured outcomes. However, we use aggregate data on school enrollment and the number of students tested to investigate whether accountability grades were significantly related to the probability that a student took the high stakes exams. In Table 1, we show there is little relationship between the percentage of students tested (in grades 3 through 8) and accountability grade. However, this section will further investigate whether grades may have affected the portion of students tested across schools.

To more formally address the possibility that improvements in test scores may have been driven in part by the composition of students being tested, we run similar regressions as above but with percentage tested as the dependent variable (Table 5). We find no significant differences in the percentage of students tested by accountability grade, either in 2007 or 2008, once we control for continuous functions of the grade inputs, and the point estimates are very small (less than 1 percentage point) and precisely estimated. Results (not reported) are very similar if we control for the percent tested in the prior school year. Thus, while we cannot rule out other forms of gaming, we do not find evidence that school receiving low accountability grades tended to exclude more students from testing in the following year.

4.5 Teacher, Parent, and Student Survey Outcomes

One of the least researched aspects of school accountability systems is the mechanisms through which they impact student achievement. Rouse et al. (2007) provide one of the first detailed analyses of this issue using surveys administered to principals before and after a change

in Florida's accountability system. We are fortunate in that the DOE surveyed teachers, parents, and students (in grades 6 and higher) as part of the new accountability system, asking them a series of questions focused on four domains: academics, safety, engagement, and communication.[21] The 2007 surveys were completed between April 30 and June 6, while the 2008 surveys were completed between March 12 and April 18; in both cases, surveys were completed after student testing in English and math but prior to the public release of test score results for the year. Surveys were confidential for parents and students and anonymous for teachers, and were collected and analyzed by an external entity contracted by the DOE.[22] All survey questions had multiple choice answers. For example, students were asked for their agreement with statements such as "my school is kept clean" on a 4 point scale ranging from "Strongly Disagree" to "Strongly Agree."

In this section, we examine the relationship between accountability grades and survey responses. We analyze teacher, parent, and student responses separately. We first examine a school's score for each of the four domains, and then proceed to examine individual questions or small groups of questions that focus on particular mechanisms through which schools behaviors may have affected student test scores.[23] First, we selected questions from all three respondent groups' surveys that described whether high expectations were set for students. There is a common belief among educators (and some work by economists, see Figlio (2005)) that setting high expectations is helpful in raising student achievement. Second, we selected questions from

[21] Information on the surveys including the complete survey instruments and guides to how they were scored can be found online at http://schools.nyc.gov/Accountability/SchoolReports/Surveys/default.htm. The "educator's guides" provided on this site contain information on the specific question items used in constructing the domain scores.
[22] Parent surveys were distributed to elementary students to be taken home, while middle and high school parents received their surveys by mail. Teacher surveys were distributed in school and students took their surveys during class time. Parents and teachers returned their completed surveys in pre-addressed, postage-paid envelopes, while schools collected student surveys. Parents and teachers also could complete their surveys online.
[23] See Appendix Table A1 for a complete accounting of where these questions are located in the environmental surveys.

all three respondent groups' surveys that described the number of classes or activities (before, after, or during school) offered in art, music, dance, theater, foreign language, and computer skills/technology. We designed this measure to address a common concern with high stakes accountability systems – that schools will shift resources to activities that increase performance on the measures that are evaluated and away from other activities that may be beneficial to students in other ways.[24]

Last, but not least, we constructed measures that were not common across all three respondent groups but that addressed particular mechanisms or issues we thought were potentially important. For parents, we examine overall satisfaction with the quality of their child's teacher and their overall satisfaction with the quality of their child's education. For teachers, we examine questions that address four issues: the extent to which administrators focus on teacher quality issues, the extent to which student achievement data is used to direct instruction, the quality of the professional development they receive, and the quality of their instructional materials. For students, we examine questions that ask for the frequency with which they are required to complete and essays or projects using "multiple sources of information" or "evidence to defend [their] opinion[s] or ideas." We originally thought these questions were linked to specific writing competencies included on state tests, but, according to our correspondence with DOE officials these questions were designed to capture whether students were given "the kind of academic work considered challenging and rigorous."

Summary statistics on survey outcomes are shown in Tables 6 and 7. Table 6 reports statistics on response rates and scores on the four domains of the survey, and Table 7 focuses on

[24] Rouse et al. (2007) find evidence that failing schools offer extra help to struggling students and lengthen instructional time. Unfortunately, these issues were not focused on in the DOE surveys. One item asked about the offering of tutoring/enrichment activities before or after school, but the wording of the question inextricably ties tutoring with enrichment; while the former typically targets struggling students, the latter might apply to all students or advanced students. Indeed, responses to this question are highly correlated with offering of foreign languages.

the subset of questions we selected for further examination. While scores in each domain and scores on each question range from 0 to 10, we normalize these variables to have a mean of 0 and standard deviation of 1 for ease of interpretation, within each school level.[25] As in Tables 1 and 2, we report averages across schools with the same accountability grade.

There are a large number of statistics in these tables; we focus on just a few stylized facts. As shown in Table 6, response rates for all three groups of respondents were fairly low in 2007, particularly among parents. Response rates grew in 2008, with somewhat larger increases among respondents associated with schools receiving low accountability grades, again, particularly among parents. For example, only about one quarter of parents in F and D schools responded in 2007, while about a third responded in schools that received higher grades. In 2008, response rates were roughly 50 percent across all schools. Response rates are important for considering how our results should be interpreted, and we return to this issue below.

Nearly all of the 2007 outcomes improve considerably and monotonically as we move from F to A schools. In most cases, F schools' scores were over 0.5 standard deviations below average, while A schools' scores tended to be about 0.3 standard deviations above average. While some of this relationship is mechanical (10 percent of the overall score determining accountability grades were based on these survey results), these results strongly suggest that survey responses must also have been highly correlated with the student achievement (levels and growth) outcomes that largely determine the accountability grade. However, the parent and

[25] There are three instances where we scale responses differently than the DOE. First, when respondents marked "don't know" or "does not apply," the DOE sometimes assigned a score signaling neutrality, but we treat these responses as missing values. Second, for questions related to offering of classes and activities in non-tested subjects (e.g., art, music), the DOE assigned 0 points when none were mentioned, 5 points if 1 or 2 classes/activities were mentioned, and 10 points if three or more were mentioned. We assign 1 point per class/activity mentioned. Third, for questions asking students about the frequency with which they are asked to do essay writing, the DOE gave 0 points for students that replied "Never," 5 points for students that replied "1 or 2 times," and 10 points for students that responded either "3 or 4 times" and "5 or more times." We make the scoring linear, assigning 0 points for "Never," 1 point for "1 or 2 times," 2 points for "3 or 4 times" and 3 points for "5 or more times."

teacher survey results for schools receiving low accountability grades improved, sometimes dramatically, between 2007 and 2008. For example, in 2007, academic scores on the parent survey averaged -0.78 standard deviations among F schools and 0.35 standard deviations among A schools but, in 2008, the average for F schools improved by 0.54 standard deviations and fell for A schools by 0.04 standard deviations. The disaggregated questions in Table 7 tell a similar story. Overall parental satisfaction with their child's education was -0.88 standard deviations for F schools and 0.32 standard deviations for A schools in 2007. In 2008, this measure for F schools increased by 0.65 standard deviations, while the A school average was virtually identical, falling by 0.001 standard deviations.

These changes suggest that the test score improvements we documented above may have been accompanied by changes on the ground that were noticeable to interested parties, particularly parents. Nevertheless, it is interesting to note that the changes in student survey outcomes are quite different than those for parents and teachers. While F schools improved relative to A schools on every measure for parents and teachers, they fell behind on all four domains and three of the four disaggregated results for students. The only student survey measure we examine on which F schools improved relative to A schools was the frequency with which students reported writing essays and doing projects, although, as discussed below, these results are reversed in our regression analysis. Parent and teacher survey responses show smaller improvements among F schools if we restrict the sample to schools with student survey data (results available upon request). Thus, part of the patterns we see in Tables 6 and 7 are due to differences between schools with younger and older students. While it would be interesting to investigate possible heterogeneity in our regression analysis below, our sample sizes (there are only 17 F schools with student survey data) are prohibitively small.

Before proceeding to the survey outcomes, we first present an analysis of response rates using the same regression specifications we used previously to examine test scores and the percent of students tested. Response rates increased the most for schools receiving lower accountability grades, particularly for parents and teachers, but it is unclear whether these increases reflect smooth relationships between response rates and the characteristics of schools receiving different grades or discontinuous relationships that would suggest the grades themselves caused increases in survey response rates. Table 8 provides estimates of the relationship between grades and response rates in both 2007 and 2008. All specifications control for continuous (quartic) functions of accountability grade inputs. As expected, there is little significant variation in response rates across accountability grades in 2007 after we control for continuous functions of the grade inputs. The coefficient on F schools for parent response rates is marginally significantly different from zero but the P-value on a test of equality with the D coefficient is 0.26. In contrast, accountability grades are significantly related to parent and teacher response rates in 2008. In particular, we see a negative relationship between response rates and accountability grades, suggesting that receipt of a low grade may have affected response rates.[26] This finding suggests that the results we present on survey outcomes should be interpreted with caution. It is generally unclear how changes in response rates might bias analysis of the survey outcomes. It may be the case that improvements within the school led to increases in test scores and survey response rates (e.g., better outreach and communication with parents). Nevertheless, it also seems reasonable that a poor grade could motivate parents and teachers who were supportive of the school but, in absence of the poor grade, would not have responded to the survey; this may be particularly true for schools near the margin of receiving

[26] For simplicity, we do not present other results here, but our findings are quite similar if we drop the 10 schools with extreme accountability scores or include controls for the 2007 response rate when examining 2008 data.

the next highest grade. We are not able to disentangle these, and other possibilities, thus, the results we present on survey outcomes should not be taken as causal.

Given the note of caution above, we proceed to present results for survey outcomes. For simplicity, we only present results for 2008, and from specifications that drop extreme accountability scores and include a quartic polynomial of the school's survey outcome in 2007. These controls do not greatly affect point estimates but noticeably reduce the standard errors. We also present specifications that control for a quartic in the school's response rate. While this is not by any means a foolproof way of removing possible endogeneity, we believe this still provides useful information on the importance of this potential bias.

Starting with scores on the four domains (Table 9), we see evidence of significant improvement on the academic, engagement, and communications scores for F and D schools among parents. These effects shrink somewhat when controls for response rate are added to the regression, but they remain statistically significant at conventional levels. Interestingly, the results for parents are not replicated among teachers and students. For teachers, the pattern of results suggest that F, D, B, and A schools all improved relative to C schools, though only the coefficients for A schools consistently come close to or meet conventional levels of statistical significance across most of the categories. Among students, we find little significant variation across accountability grades, though with the reduced number of schools in the student regressions, the standard errors are considerably larger. The point estimates suggest that, if anything, both F and A schools performed poorly on the student surveys relative to schools with less extreme grades.

We next turn to the two sets of academic questions that provide common measures across all three subgroups: whether the school holds high expectations for students and the extent of

course offerings in non-tested subject areas. The results for expectations (Columns 1 and 2 of Table 10) generally mirror those for the overall score for academics shown in Table 9, though controlling for response rates generally reduces the coefficients to the point where there are no longer statistically significant. Interestingly, survey results from all groups suggest that, if anything, schools with lower accountability grades had greater offerings of non-tested courses. The inclusion of response rate as a control has little influence on these results. If we limit the course offering measures to examine courses offered during the regular school day, the results are quite similar. Of course, differential response rates by parents, teachers, and students who are more familiar with these programs could drive these results. Given this endogeneity problem, we would be reluctant to conclude that schools receiving low grades responded by shifting resources into courses like art, music, and dance. However, if schools with low grades were indeed cutting back on these courses, the bias from endogeneity would need to be substantial.

Turning to the most specific survey questions (Table 11), we find very large and significant effects of F and D grades on parental evaluations of teacher quality and their children's overall education. These effects remain significant when we control for response rates. We also find some, albeit weak, evidence that teachers in F and D schools placed greater emphasis on using student achievement data to make instructional decisions and teachers in D schools were given higher quality instructional materials. We find stronger, though somewhat surprising evidence that teachers in schools receiving lower accountability grades felt there was *less* focus on teaching quality by school leaders (i.e., classroom visits, feedback, and priority placed on the quality of teaching). Last, we do find evidence that students in schools receiving F and D grades spent less time doing work that involved essays and projects. One story which reconciles all of these findings is that principals in F and D schools placed greater weight on

direct instruction of skills that would lead to improvements on the state examinations, which involved a greater use of data driven instruction and less of what teachers deem a focus on teaching quality. In the end, however, parents seem to be quite satisfied with the results.

5. Conclusion

The results of our analysis suggest that the new accountability system put in place in New York City had important effects in the months that followed its launch in the fall of 2007. Schools that received very low accountability grades (F or D) saw improved test scores in math and English. For example, we estimate that the impact of attending an F school (as compared to a C school) on students' math test scores was roughly 0.1 student-level standard deviations. Our use of the discontinuous assignment of accountability grades supports the notion that our analysis provides causal estimates of the impact of accountability on student academic achievement and is not confounded by a spurious relationship between grade assignment and other factors affecting student performance.

In order to provide additional insight into the impact of accountability pressure on low performing schools, we examine a set of complementary outcomes using surveys of parents, teachers, and students. Because response rates on these surveys were less than complete, there is the possibility that endogeneity of response affects the results of this secondary analysis. However, given this caveat, we find fairly strong evidence that satisfaction with academic quality rose significantly for parents of children in F and D schools. Our analysis also provides suggestive evidence that schools may have achieved this complementary outcome, and the test score improvements, through greater use of student achievement data and direct instruction. However, we do not find evidence that these schools cut back on offering instruction in non-tested subjects such as art and music.

These results suggest several avenues for future research. While evidence that accountability pressure can induce improvements in student achievement has grown, we still know relatively little about the pathways through which accountability systems can achieve these outcomes. We offer some insights into this issue through our examination of survey data, but further research is needed on the actions taken by principals, teachers, and others in response to the rewards and consequences presented by accountability systems. Moreover, many questions remain regarding variation in the type and severity of accountability incentives. Is the stigma of an "F" or the possibility of being fired the crucial factor in motivating principals of poor performing schools? Are financial bonuses more effective if paid to principals, teachers, or the students themselves? More research is needed on these and other questions regarding behavioral responses to incentives generated by accountability.

References

Alderman, D. and Powers, D. (1989), "The Effects of Special Preparation on SAT-Verbal Scores," American Educational Research Journal, 17(2): 239-251.

Black, S. (1999) "Do Better Schools Matter? Parental Valuation of Elementary Education," Quarterly Journal of Economics 114(2): 577-599

Chakrabarti, R. (2006) "Vouchers, Public School Response and the Role of Incentives: Evidence from Florida," Working Paper, Federal Reserve Bank of New York.

Chakrabarti, R. (2008) "Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs," Federal Reserve Bank of New York Staff Reports #315.

Chay, K. McEwan, P. and Urquiola, M. (2005) "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools," American Economic Review, 95(4): 1237-1258.

Chiang, H. (2007) "How Accountability Pressure on Failing Schools Affects Student Achievement," Unpublished Manuscript Harvard University.

Cullen, J.B. and Reback, R. (2006) "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System," In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Fan, Jianqing and Irene Gijbels (1997) "Local Polynomial Modeling and its Applications," Chapman and Hall: London.

Figlio, D. and Lucas, M. (2004) "What's in a Grade? School Report Cards and the Housing Market," American Economic Review, 94(3): 591-604

Figlio, D. (2005) "Names, Expectations and the Black-White Test Score Gap," NBER working paper #11195.

Figlio, D. (2006) "Testing, Crime, and Punishment," Journal of Public Economics 90, 837-851.

Figlio, D. and Getzler, L. (2006) "Accountability, Ability, and Disability: Gaming the System?" In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Figlio, D. and Rouse, C. (2006) "Do Accountability and Voucher Threats Improve Low-Performing Schools?" Journal of Public Economics 90, 239-255.

Figlio, D. and Winicki, J. (2005) "Food for Thought? The Effects of School Accountability Plans on School Nutrition," Journal of Public Economics 89, 381-394.

Gootman, E. and Medina, J. “50 City Schools Get Failing Grade in a New System,” New York Times, November 6, 2007.

Gootman, E. “The Day After School Grades Come In, Parents Are Buzzing,” New York Times, November 7, 2007.

Hastings, J., Kane, J. and Staiger, D. (2007), “Preferences and Heterogeneous Treatment Effects in a Public School Choice Lottery,” NBER Working Paper 12145.

Hastings, J. and Weinstein, J. (2008) “Information, School Choice, and Academic Achievement: Evidence from Two Experiments,” Quarterly Journal of Economics, 123(4): 1373- 1414.

Hanushek, E., Kain, J. and Rivkin, S. (2004) “Disruption Versus Tiebout Improvement: The Costs and Benefits of Switching Schools,” Journal of Public Economics, 88(9): 1721-1746.

Hanushek, E. and Raymond, M. (2005) “Does School Accountability Lead to Improved School Performance?” Journal of Policy Analysis and Management 24(2), 297 – 327.

Hoxby, C. and Weingarth-Salyer, G. (2005), “Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects,” Unpublished Manuscript, Harvard University.

Jacob, B. (2005) “Accountability, Incentives and Behavior: the Impact of High-stakes Testing in the Chicago Public Schools,” Journal of Public Economics 89(5-6), 761-796.

Jacob, B, and Lefgren, L (2004) “Remedial Education and Student Achievement: A Regression-Discontinuity Analysis,” Review of Economics and Statistics 86(1): 226-244.

Jacob, B. and Levitt, S. (2003) “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating,” Quarterly Journal of Economics 118(3), 843-877.

Kane, T. and Staiger, D. (2002) “The Promise and Pitfalls of Using Imprecise School Accountability Measures,” Journal of Economic Perspectives 16(4): 91–114

Kane, T., Rockoff, J. and Staiger, D. (2008) “What Does Certification Tell us about Teacher Effectiveness? Evidence from New York City,” Economics of Education Review, 27(6): 615-631.

Krieg, J. (2008) “Are students left behind? The Distributional Effects of No Child Left Behind,” Education Finance and Policy 3(2), 250-281.

Ladd, H. and Zelli, A. (2002) “School-based accountability in North Carolina: The responses of school principals,” Educational Administration Quarterly 38(4), 494-529.

Mizala, A. and Urquiola, M. (2008) “School markets: The impact of information approximating schools’ effectiveness,” Unpublished Manuscript.

Neal, D. and Whitmore Schanzenbach, D. (2007) “Left behind by design: Proficiency counts and test-based accountability,” NBER working paper #13293.

Medina, J. and Gootman, E. “Schools Brace to be Scored on a Scale of A to F,” New York Times, November 4, 2007.

Peterson, P. and West, M., eds. No Child Left Behind?: The Politics and Practice of School Accountability Washington D.C.: Brookings Institution Press, 2003

Reback, R. (forthcoming) “Teaching to the Rating: School Accountability and the Distribution of Student Achievement” Journal of Public Economics.

Rouse, C., Hannaway, J., Goldhaber, D., and Figlio, D. (2007) “Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure,” NBER working paper #13681.

Van der Klaauw, W. (2002) “Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–Discontinuity Approach,” International Economic Review 43(4): 1249-1287

Table 1: Descriptive Statistics by Accountability Grade

	Progress Report Grade				
	F	D	C	B	A
Number of Schools	42	87	259	371	226
Type of School					
Elementary	59.5%	58.6%	59.8%	60.1%	56.2%
K-8	11.9%	9.2%	12.0%	12.4%	11.5%
Middle	28.6%	32.2%	28.2%	27.5%	32.3%
Enrollment	518	622	681	717	635
% of Enrollment in Grades 3-8	56.4%	61.5%	64.5%	64.4%	64.0%
NCLB Status					
Restructuring	11.9%	16.1%	15.1%	14.4%	9.3%
Needs Improvement	16.7%	14.9%	14.7%	16.3%	11.6%
In Good Standing	71.4%	69.0%	70.2%	69.4%	79.1%
Student Characteristics					
% Free Lunch	76.7%	77.7%	67.2%	68.9%	67.6%
% Special Education	10.9%	10.3%	10.4%	8.5%	7.7%
% English Language Learner	9.8%	11.6%	11.4%	13.0%	12.6%
% Black	44.9%	44.6%	37.9%	32.4%	27.3%
% Hispanic	40.7%	40.8%	37.3%	39.9%	42.9%
% White	9.7%	9.6%	15.1%	14.2%	13.1%
% Asian	4.1%	4.4%	9.1%	13.0%	16.1%
Test Score Outcomes '06-'07					
Average Scale Score English	641.8	644.7	650.8	654.3	659.7
Above Median English Score	23.8%	25.3%	44.0%	53.1%	66.4%
% Students Tested in English (Grades 3 - 8)	93.7%	95.1%	94.4%	95.0%	94.7%
Average Scale Score Math	653.5	657.3	665.3	670.5	677.2
Above Median Math Score	19.0%	23.0%	42.5%	54.4%	67.7%
% Students Tested in Math (Grades 3 - 8)	94.5%	96.3%	96.0%	96.6%	96.3%
Test Score Outcomes '07-'08					
Average Scale Score English	648.4	648.8	654.1	656.9	661.2
Above Median English Score	23.8%	28.7%	45.2%	54.2%	61.9%
% Students Tested in English (Grades 3 - 8)	97.0%	97.1%	96.9%	97.2%	97.5%
Average Scale Score Math	662.0	664.2	669.9	675.3	681.1
Above Median Math Score	23.8%	25.3%	41.3%	53.9%	68.1%
% Students Tested in Math (Grades 3 - 8)	98.0%	98.3%	98.4%	98.8%	98.9%
Progress Report Scores					
Overall Score	23.0	35.0	44.9	56.5	72.1
Environment Score	4.9	5.7	6.8	7.9	9.1
Performance Score	10.2	12.0	14.7	16.8	20.4
Progress Score	7.6	16.7	22.2	29.4	38.2
Additional Credit	0.3	0.7	1.2	2.4	4.3
Peer Index (mean = 0, s.d. = 1)	-0.389	-0.299	-0.031	0.044	0.151
Quality Review Rating					
Undeveloped	16.7%	14.9%	8.5%	6.7%	2.2%
Proficient	66.7%	67.8%	56.0%	53.4%	50.0%
Well Developed	16.7%	17.2%	35.1%	39.9%	47.8%

Table 2: Descriptive Statistics by NCLB Status

	NCLB Status		
	In/Planning Restructuring	Needs Improvement	In Good Standing
Number of Schools	132	144	705
Type of School			
Elementary	35.6%	46.5%	66.2%
K-8	8.3%	5.6%	13.8%
Middle	56.1%	47.9%	20.0%
Enrollment	862	812	610
% of Enrollment in Grades 3-8	80.7%	73.7%	56.7%
Student Characteristics			
% Free Lunch	81.6%	75.4%	65.6%
% Special Education	11.2%	10.6%	8.4%
% English Language Learner	19.6%	15.6%	10.0%
% Black	32.2%	32.9%	35.0%
% Hispanic	57.8%	47.7%	34.9%
% White	3.3%	10.3%	16.3%
% Asian	6.0%	8.5%	13.3%
Test Score Outcomes '06-'07			
Average Scale Score English	637.3	644.9	658.2
% Students at Level 4 English	1.2%	2.7%	5.9%
% Students Tested in English (Grades 3 - 8)	93.4%	94.2%	95.2%
Average Scale Score Math	649.7	659.6	674.5
% Students at Level 4 Math	8.2%	14.1%	24.2%
% Students Tested in Math (Grades 3 - 8)	95.9%	96.2%	96.3%
Progress Report Scores			
Overall Score	52.3	52.0	54.3
Environment Score	6.4	6.6	8.0
Performance Score	13.7	15.0	17.1
Progress Score	29.2	27.6	27.1
Additional Credit	3.1	2.7	2.1
Peer Index *(mean = 0, s.d. = 1)*	-0.637	-0.334	0.201
Quality Review Rating			
Undeveloped	18.2%	13.9%	4.0%
Proficient	62.1%	55.6%	53.6%
Well Developed	18.9%	30.6%	42.4%

Table 3: The Impact of Accountability Grades on Achievement

	2007 *(Placebo)*		2008					
	Math	English	Math			English		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accountability Grade								
F	1.672	0.118	5.745	4.412	4.260	2.682	2.565	2.177
	(1.177)	(0.965)	(1.556)**	(1.216)**	(1.241)**	(1.292)*	(1.142)*	(1.147)+
D	0.884	0.206	2.873	2.133	2.131	0.657	0.449	0.386
	(0.607)	(0.521)	(0.823)**	(0.686)**	(0.695)**	(0.630)	(0.547)	(0.549)
B	-0.167	0.392	-0.160	-0.122	-0.145	-0.420	-0.621	-0.605
	(0.684)	(0.500)	(0.845)	(0.625)	(0.629)	(0.555)	(0.478)	(0.475)
A	-0.341	1.282	-1.603	-1.414	-1.433	-0.792	-1.545	-1.390
	(1.218)	(0.903)	(1.557)	(1.157)	(1.170)	(1.011)	(0.906)+	(0.889)
Test that D = F (p-value)	0.38	0.91	0.03	0.04	0.05	0.08	0.04	0.07
Test that A = B (p-value)	0.81	0.11	0.14	0.07	0.08	0.53	0.09	0.14
Report Element Scores & Peer Index (Quartic)	√	√	√	√	√	√	√	√
Prior Scale Score (Quartic)				√	√		√	√
Dropped Extreme A and F Schools					√			√
Observations	985	985	985	985	975	985	985	975

Notes: Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%. Specifications with category scores and peer index also include controls for school levels and interactions between school levels and report element scores and peer index.

Table 4: Heterogeneity of Results Across Schools

	Math						English					
Interaction Term	Elem/K-8 Schools	Middle Schools	In Good Standing for NCLB		Prior Scale Score Above Median		Elem/K-8 Schools	Middle Schools	In Good Standing for NCLB		Prior Scale Score Above Median	
			No	Yes	No	Yes			No	Yes	No	Yes
Accountability Grade												
F	3.219	6.394	4.213	4.494	4.214	4.169	2.065	2.694	2.604	2.040	3.153	0.531
	(1.364)*	(2.779)*	(1.694)*	(1.386)**	(1.424)**	(1.760)*	(1.369)	(1.997)	(1.108)*	(1.342)	(1.201)**	(1.816)
D	1.312	3.330	1.650	2.494	1.973	2.312	0.457	0.285	-0.221	0.884	0.705	0.550
	(0.796)+	(1.433)*	(0.962)	(0.865)**	(0.809)*	(1.113)*	(0.679)	(0.943)	(0.745)	(0.665)	(0.650)	(0.766)
B	0.212	-0.814	-0.708	-0.038	0.033	-0.262	-0.594	-0.618	-0.288	-0.812	0.004	-1.265
	(0.719)	(1.189)	(0.787)	(0.736)	(0.752)	(0.752)	(0.570)	(0.836)	(0.632)	(0.512)	(0.543)	(0.535)*
A	-0.293	-3.110	-1.892	-1.418	-1.257	-1.446	-1.387	-1.348	-1.097	-1.509	-0.573	-2.014
	(1.320)	(2.202)	(1.326)	(1.243)	(1.332)	(1.248)	(1.093)	(1.546)	(1.039)	(0.923)	(1.001)	(0.908)*
F test: D = F (p-value)	0.12	0.18	0.12	0.12	0.08	0.31	0.18	0.18	0.01	0.35	0.02	0.99
F test: A = B (p-value)	0.52	0.11	0.23	0.07	0.17	0.14	0.24	0.44	0.25	0.25	0.42	0.18
F test: Same effect for 2 groups (p-value)												
F vs. D	0.65		0.77		0.86		0.71		0.24		0.22	
D	0.22		0.47		0.79		0.88		0.21		0.86	
B	0.46		0.44		0.72		0.98		0.39		0.02	
A vs. B	0.28		0.84		0.91		0.96		0.87		0.80	
Report Elements & Peer Index (Quartic)	√		√		√		√		√		√	
Prior Scale Score (Quartic)	√		√		√		√		√		√	
Dropped Extreme Scores	√		√		√		√		√		√	
Observations	975		975		975		975		975		975	

Notes: Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%. Dotted lines separate estimates from each of six regressions where the coefficients on accountability grades are allowed to differ across types of schools. All specifications include controls for school levels and interactions between school levels and report elements and peer index

Table 5: School Accountability and Percentage of Students Tested

	Math		English	
	2007	2008	2007	2008
Accountability Grade				
F	0.007	0.002	-0.000	-0.001
	(0.008)	(0.007)	(0.009)	(0.007)
D	0.006	-0.001	-0.000	-0.001
	(0.005)	(0.004)	(0.005)	(0.004)
B	0.003	-0.001	0.005	0.003
	(0.004)	(0.004)	(0.004)	(0.003)
A	-0.005	-0.002	0.000	0.009
	(0.007)	(0.006)	(0.007)	(0.006)
Category Scores & Peer Index (Quartic)	√	√	√	√
Observations	985	985	985	985

Notes: Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%. Percent of students tested is defined by dividing the total tested by enrollment in grades three to eight as measured in October 31st, 2006 (for 2007) and November 5, 2007 (for 2008). Specifications with category scores and peer index also include controls for school levels and interactions between school levels and category scores and peer index.

Table 6: Descriptive Statistics on Environment Survey Response Rates and Scores

	Progress Report Grade				
	F	D	C	B	A
Number of Schools	42	87	259	371	226
Parent Survey Results, 2007					
Response Rate	24.5%	25.3%	30.1%	31.2%	34.0%
Academic	-0.775	-0.359	-0.131	0.046	0.349
Engagement	-0.336	-0.182	-0.110	-0.009	0.268
Communication	-0.732	-0.385	-0.151	0.034	0.400
Safety	-0.657	-0.602	-0.144	0.057	0.429
Change in Parent Results 2007 to 2008					
Response Rate	24.5%	28.1%	21.4%	18.1%	15.3%
Academic	0.536	0.234	-0.065	-0.037	-0.044
Engagement	0.098	-0.003	-0.060	0.002	0.053
Communication	0.603	0.355	-0.070	-0.041	-0.098
Safety	0.383	0.272	-0.100	-0.039	0.002
Teacher Survey Results, 2007					
Response Rate	46.4%	42.7%	46.1%	46.3%	49.1%
Academic	-0.965	-0.538	-0.162	0.122	0.373
Engagement	-0.781	-0.438	-0.187	0.126	0.319
Communication	-0.775	-0.411	-0.136	0.106	0.285
Safety	-0.849	-0.610	-0.192	0.094	0.460
Change in Teacher Results, 2007 to 2008					
Response Rate	20.7%	29.8%	19.7%	16.7%	16.2%
Academic	0.334	0.302	-0.019	-0.078	-0.033
Engagement	0.352	0.272	0.039	-0.093	-0.065
Communication	0.325	0.296	-0.006	-0.058	-0.077
Safety	0.404	0.212	-0.009	-0.055	-0.058
Student Survey Results, 2007					
Response Rate	60.7%	71.6%	71.7%	73.4%	77.7%
Academic	-0.706	-0.264	-0.180	0.137	0.137
Engagement	-0.654	-0.499	-0.255	0.141	0.281
Communication	-0.312	-0.400	-0.268	0.129	0.224
Safety	-0.653	-0.510	-0.308	0.046	0.460
Change in Student Results, 2007 to 2008					
Response Rate	11.3%	14.4%	13.2%	15.1%	11.4%
Academic	-0.163	-0.135	0.009	-0.066	0.176
Engagement	-0.018	0.210	0.061	-0.106	0.050
Communication	0.017	0.133	0.107	-0.112	0.041
Safety	-0.076	0.071	0.013	0.003	-0.003

Note: The number of schools with student survey data differs from the total number of schools, as only students that were in 6th or a higher grade were surveyed. The number of schools with student survey data is as follows: 2007 -- 17 (F), 35 (D), 116 (C), 182 (B), 129 (A); 2008 -- 15 (F), 36 (D), 115 (C), 177 (B), 117 (A). 22 schools had students in grades that were surveyed in 2007 and no longer had students in these grades in 2008, while two schools had no students in grades that were surveyed in 2007 and did have students in these grades in 2008. There were nine schools in 2007 that had students in surveyed grades but zero student respondents and eight such schools in 2008.

Table 7: Discriptive Statistics on Specific Outcomes from Environment Survey

	Progress Report Grade				
	F	D	C	B	A
Number of Schools	42	87	259	371	226
Parent Survey Results, 2007					
Course Offerings	-0.357	-0.340	-0.062	0.000	0.275
High Expectations	-0.644	-0.288	-0.124	0.036	0.312
Teacher Quality	-0.649	-0.372	-0.109	0.039	0.323
Overall Satisfaction with Education	-0.880	-0.356	-0.122	0.070	0.322
Change in Parent Results 2007 to 2008					
Course Offerings	0.120	0.120	0.018	-0.014	-0.070
High Expectations	0.316	0.126	-0.080	-0.022	0.019
Teacher Quality	0.620	0.262	-0.021	-0.045	-0.114
Overall Satisfaction with Education	0.652	0.233	-0.078	-0.069	-0.001
Teacher Survey Results, 2007					
Course Offerings	-0.527	-0.266	-0.059	0.089	0.124
High Expectations	-0.892	-0.370	-0.150	0.090	0.333
Focus on Teaching by School Leaders	-0.863	-0.304	-0.157	0.137	0.233
Use of Student Data	-0.859	-0.189	-0.156	0.130	0.196
Professional Development Quality	-0.655	-0.167	-0.154	0.115	0.171
Quality of Instructional Materials	-0.659	-0.278	-0.096	0.066	0.230
Change in Teacher Results, 2007 to 2008					
Course Offerings	0.165	0.197	-0.037	-0.064	0.044
High Expectations	0.172	0.065	-0.039	-0.048	0.065
Focus on Teaching by School Leaders	0.261	0.118	-0.015	-0.038	-0.017
Use of Student Data	0.492	0.186	0.022	-0.090	-0.044
Professional Development Quality	0.378	0.142	0.079	-0.116	-0.035
Quality of Instructional Materials	0.108	0.084	-0.051	-0.059	0.096
Student Survey Results, 2007					
Course Offerings	-0.286	-0.214	-0.096	0.098	0.066
High Expectations	-0.682	-0.562	-0.302	0.154	0.304
Essays and Projects	-0.882	-0.276	-0.156	0.151	0.114
Group and Hands-on Learning Activities	-0.428	-0.445	-0.295	0.139	0.262
Change in Student Results, 2007 to 2008					
Course Offerings	-0.060	0.002	0.031	-0.062	0.068
High Expectations	-0.141	0.194	0.092	-0.071	0.010
Essays and Projects	0.232	-0.087	0.095	-0.048	-0.002
Group and Hands-on Learning Activities	0.020	0.063	0.177	-0.106	-0.016

Note: The number of schools with student survey data differs from the total number of schools, as only students that were in 6th or a higher grade were surveyed. The number of schools with student survey data is as follows: 2007– 17 (F), 35 (D), 116 (C), 182 (B), 129 (A); 2008 – 15 (F), 36 (D), 115 (C), 177 (B), 117 (A). 22 schools had students in grades that were surveyed in 2007 and no longer had students in these grades in 2008, while two schools had no students in grades that were surveyed in 2007 and did have students in these grades in 2008. There were nine schools in 2007 that had students in surveyed grades but zero student respondents and eight such schools in 2008.

Table 8: School Accountability and Survey Response Rates

	2007			2008		
	Parents	Teachers	Students	Parents	Teachers	Students
Accountability Grade	(1)	(2)	(3)	(4)	(5)	(6)
F	-0.039	0.008	-0.038	0.063	0.057	-0.044
	(0.023)+	(0.043)	(0.079)	(0.056)	(0.045)	(0.045)
D	-0.018	-0.041	0.008	0.074	0.055	-0.011
	(0.016)	(0.031)	(0.037)	(0.037)*	(0.026)*	(0.026)
B	0.005	-0.021	-0.021	-0.060	-0.040	0.009
	(0.015)	(0.023)	(0.030)	(0.029)*	(0.023)+	(0.020)
A	0.019	-0.016	0.036	-0.110	-0.070	0.027
	(0.028)	(0.041)	(0.059)	(0.057)+	(0.044)	(0.034)
Report Elements & Peer Index (Quartic)	√	√	√	√	√	√
Observations	985	974	479	985	985	460

Notes: Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%. Specifications with category scores and peer index also include controls for school levels and interactions between school levels and report elements and peer index.

Table 9: Survey Outcomes (Four Domain Scores)

Panel A: Parent Survey Results	Academics		Engagement		Communication		Safety	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accountability Grade								
F	0.562	0.441	0.313	0.246	0.486	0.369	0.229	0.145
	(0.191)**	(0.185)*	(0.228)	(0.222)	(0.217)*	(0.200)+	(0.184)	(0.184)
D	0.349	0.232	0.285	0.215	0.402	0.288	0.228	0.137
	(0.123)**	(0.115)*	(0.122)*	(0.119)+	(0.123)**	(0.113)*	(0.115)*	(0.105)
B	-0.138	-0.062	-0.119	-0.065	-0.024	0.052	0.032	0.101
	(0.096)	(0.087)	(0.089)	(0.084)	(0.092)	(0.083)	(0.077)	(0.070)
A	-0.326	-0.189	-0.183	-0.093	-0.063	0.071	0.125	0.244
	(0.188)+	(0.170)	(0.170)	(0.159)	(0.181)	(0.161)	(0.157)	(0.140)+
Test that D = F (p-value)	0.22	0.22	0.90	0.88	0.67	0.66	1.00	0.96
Test that A = B (p-value)	0.13	0.26	0.55	0.78	0.74	0.85	0.37	0.13
Observations	975	975	975	975	975	975	975	975
Panel B: Teacher Survey Results	Academics		Engagement		Communication		Safety	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accountability Grade								
F	0.144	0.104	0.114	0.078	0.078	0.023	0.307	0.290
	(0.179)	(0.174)	(0.187)	(0.179)	(0.196)	(0.185)	(0.160)+	(0.159)+
D	0.224	0.166	0.162	0.104	0.189	0.131	0.142	0.098
	(0.121)+	(0.117)	(0.131)	(0.126)	(0.131)	(0.125)	(0.109)	(0.106)
B	0.076	0.110	0.092	0.126	0.148	0.182	0.063	0.088
	(0.087)	(0.085)	(0.092)	(0.090)	(0.092)	(0.090)*	(0.079)	(0.077)
A	0.216	0.267	0.271	0.319	0.298	0.358	0.194	0.237
	(0.160)	(0.158)+	(0.171)	(0.169)+	(0.167)+	(0.162)*	(0.150)	(0.149)
Test that D = F (p-value)	0.66	0.72	0.79	0.88	0.57	0.56	0.29	0.21
Test that A = B (p-value)	0.17	0.12	0.11	0.08	0.15	0.09	0.17	0.12
Observations	963	963	963	963	963	963	963	963
Panel C: Student Survey Results	Academics		Engagement		Communication		Safety	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
F	-0.582	-0.424	-0.174	-0.080	-0.605	-0.554	-0.018	0.008
	(0.414)	(0.389)	(0.322)	(0.319)	(0.310)+	(0.305)+	(0.307)	(0.298)
D	-0.164	-0.128	0.009	0.026	-0.120	-0.107	-0.003	-0.009
	(0.234)	(0.219)	(0.165)	(0.156)	(0.166)	(0.164)	(0.200)	(0.187)
B	0.132	0.085	0.025	-0.003	-0.011	-0.019	-0.063	-0.080
	(0.168)	(0.167)	(0.139)	(0.143)	(0.156)	(0.159)	(0.108)	(0.110)
A	-0.108	-0.176	-0.197	-0.235	-0.415	-0.435	-0.265	-0.277
	(0.289)	(0.288)	(0.263)	(0.266)	(0.272)	(0.276)	(0.215)	(0.217)
Test that D = F (p-value)	0.21	0.35	0.48	0.68	0.03	0.05	0.94	0.94
Test that A = B (p-value)	0.15	0.12	0.17	0.15	0.01	0.01	0.16	0.16
Observations	444	444	444	444	444	444	444	444
Report Elements & Peer Index (Quartic)	√	√	√	√	√	√	√	√
Prior Domain Score (Quartic)	√	√	√	√	√	√	√	√
Dropped Extreme Scores	√	√	√	√	√	√	√	√
Response Rate (Quartic)		√		√		√		√

Notes: All dependent variables have been standardized to have mean zero and standard deviation one. Specifications with report element scores and peer index also include controls for school levels and interactions between school levels and report elements and peer index. Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%

Table 10: Survey Outcomes (High Expectations and Course Offerings)

Panel A: Parent Survey Results	High Expectations		Course Offerings	
	(1)	(2)	(5)	(6)
Accountability Grade				
F	0.359	0.271	0.173	0.167
	(0.192)+	(0.196)	(0.145)	(0.144)
D	0.241	0.161	0.194	0.194
	(0.111)*	(0.105)	(0.120)	(0.123)
B	-0.022	0.036	-0.118	-0.125
	(0.088)	(0.084)	(0.072)	(0.071)+
A	-0.047	0.053	-0.266	-0.277
	(0.179)	(0.165)	(0.148)+	(0.147)+
Test that D = F (p-value)	0.50	0.54	0.89	0.85
Test that A = B (p-value)	0.83	0.88	0.12	0.12
Observations	975	975	975	975
Panel B: Teacher Survey Results	**High Expectations**		**Course Offerings**	
	(1)	(2)	(5)	(6)
Accountability Grade				
F	0.074	0.026	0.202	0.182
	(0.155)	(0.156)	(0.243)	(0.240)
D	0.192	0.134	0.256	0.242
	(0.106)+	(0.103)	(0.128)*	(0.131)+
B	0.051	0.087	-0.037	-0.027
	(0.082)	(0.079)	(0.097)	(0.096)
A	0.155	0.215	0.002	0.017
	(0.152)	(0.148)	(0.176)	(0.175)
Test that D = F (p-value)	0.43	0.46	0.81	0.79
Test that A = B (p-value)	0.28	0.17	0.70	0.67
Observations	975	975	975	975
Panel C: Student Survey Results	**High Expectations**		**Course Offerings**	
	(1)	(2)	(5)	(6)
Accountability Grade				
F	-0.342	-0.320	0.092	0.079
	(0.349)	(0.345)	(0.378)	(0.381)
D	0.023	0.022	0.078	0.072
	(0.192)	(0.187)	(0.181)	(0.178)
B	-0.014	-0.021	-0.017	-0.017
	(0.135)	(0.138)	(0.143)	(0.145)
A	-0.412	-0.416	-0.100	-0.094
	(0.245)+	(0.250)+	(0.265)	(0.265)
Test that D = F (p-value)	0.20	0.23	0.96	0.98
Test that A = B (p-value)	0.01	0.01	0.63	0.66
Observations	444	444	444	444
Report Elements & Peer Index (Quartic)	√	√	√	√
Prior Survey Score (Quartic)	√	√	√	√
Dropped Extreme Scores	√	√	√	√
Response Rate (Quartic)		√		√

Notes: All dependent variables have been standardized to have mean zero and standard deviation one. Robust standard errors in parentheses. Specifications with report element scores and peer index also include controls for school levels and interactions between school levels and report element scores and peer index.
+ significant at 10%; * significant at 5%; ** significant at 1%

Table 11: Survey Outcomes (Specific Academic Survey Questions)

Survey Response Group:	Parents				Teachers								Students			
Survey Question(s):	Teacher Quality		Overall Education Quality		Use of Student Data		Focus on Teaching by School Leaders		Quality of Professional Development		Quality of Instructional Materials		Frequency of Essays/Projects		Group and Hands-on Learning Activities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Accountability Grade																
F	0.641	0.532	0.626	0.522	0.284	0.224	-0.048	-0.094	0.139	0.101	0.110	0.059	-0.519	-0.523	-0.604	-0.679
	(0.220)**	(0.214)*	(0.176)**	(0.169)**	(0.225)	(0.219)	(0.197)	(0.190)	(0.215)	(0.208)	(0.184)	(0.183)	(0.325)	(0.324)	(0.327)+	(0.339)*
D	0.257	0.149	0.365	0.259	0.277	0.213	0.160	0.093	0.206	0.141	0.284	0.244	-0.377	-0.380	-0.313	-0.334
	(0.121)*	(0.114)	(0.116)**	(0.107)*	(0.145)+	(0.140)	(0.122)	(0.116)	(0.145)	(0.140)	(0.137)*	(0.135)+	(0.182)*	(0.181)*	(0.185)+	(0.190)+
B	-0.118	-0.046	-0.106	-0.030	0.026	0.071	0.137	0.180	0.024	0.061	-0.050	-0.024	0.012	0.013	-0.155	-0.128
	(0.103)	(0.096)	(0.091)	(0.084)	(0.099)	(0.097)	(0.092)	(0.088)*	(0.099)	(0.098)	(0.096)	(0.096)	(0.149)	(0.150)	(0.169)	(0.172)
A	-0.276	-0.145	-0.194	-0.062	0.071	0.146	0.309	0.386	0.154	0.212	0.106	0.149	-0.129	-0.125	-0.447	-0.427
	(0.198)	(0.183)	(0.177)	(0.161)	(0.183)	(0.181)	(0.167)+	(0.161)*	(0.188)	(0.187)	(0.179)	(0.180)	(0.256)	(0.257)	(0.286)	(0.285)
Test that D = F (p-value)	0.36	0.45	0.22	0.18	0.98	0.96	0.28	0.30	0.34	0.38	0.33	0.29	0.54	0.54	0.46	0.47
Test that A = B (p-value)	0.05	0.05	0.88	1.52	0.70	0.52	0.11	0.05	0.74	0.84	0.16	0.12	0.36	0.37	0.24	0.18
Reponse Rate (Quartic)		√		√		√		√		√		√		√		√
Observations	975	975	975	975	975	975	975	975	975	975	975	975	444	444	444	444

Notes: All dependent variables have been standardized to have mean zero and standard deviation one. Robust standard errors in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%. All specifications drop schools with extreme scores and include a quartic in category scores and school peer index, controls for school levels and interactions between school levels and category scores and peer index, and a quartic in prior survey scores.

Figure 1: Accountability Grade Plotted Against Final Score, by Type of School

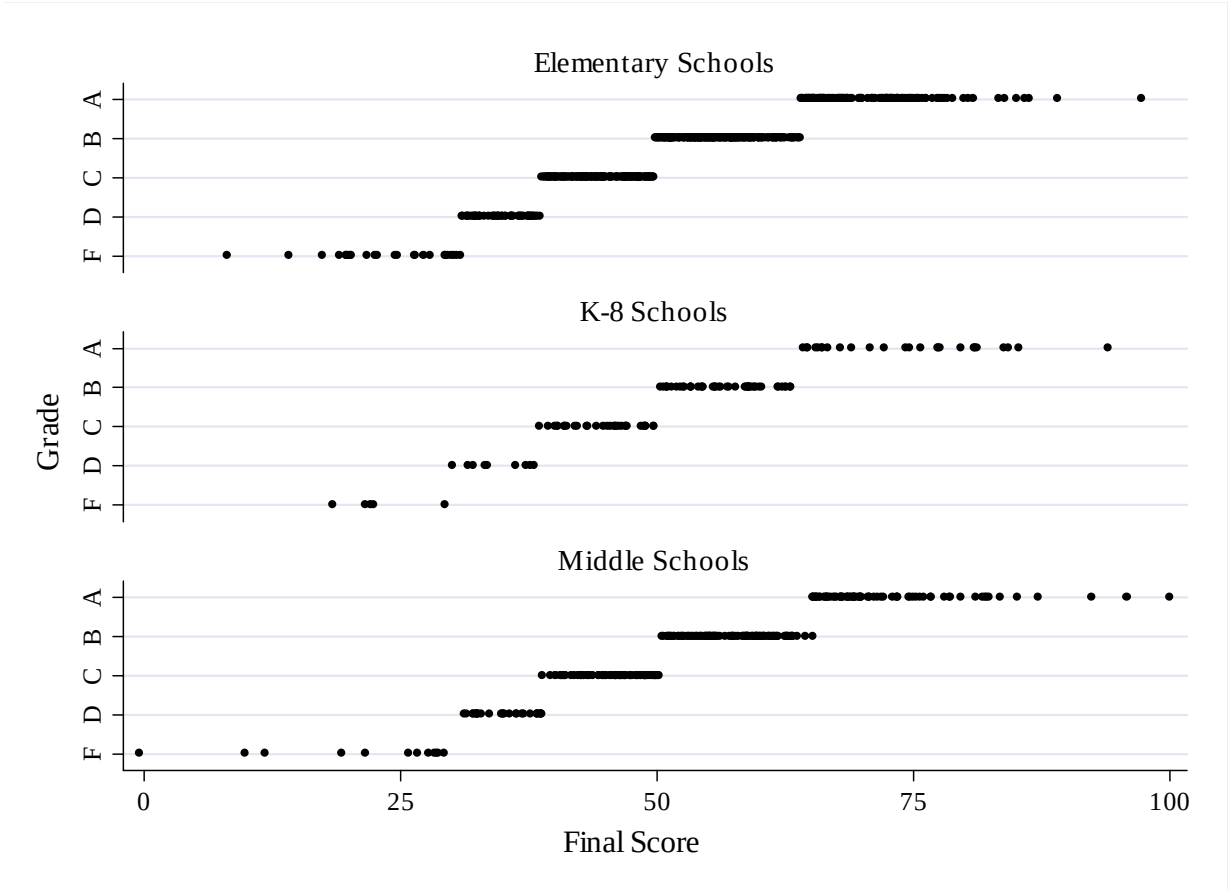


Figure 2: Timeline of Events Related to Accountability Implementation

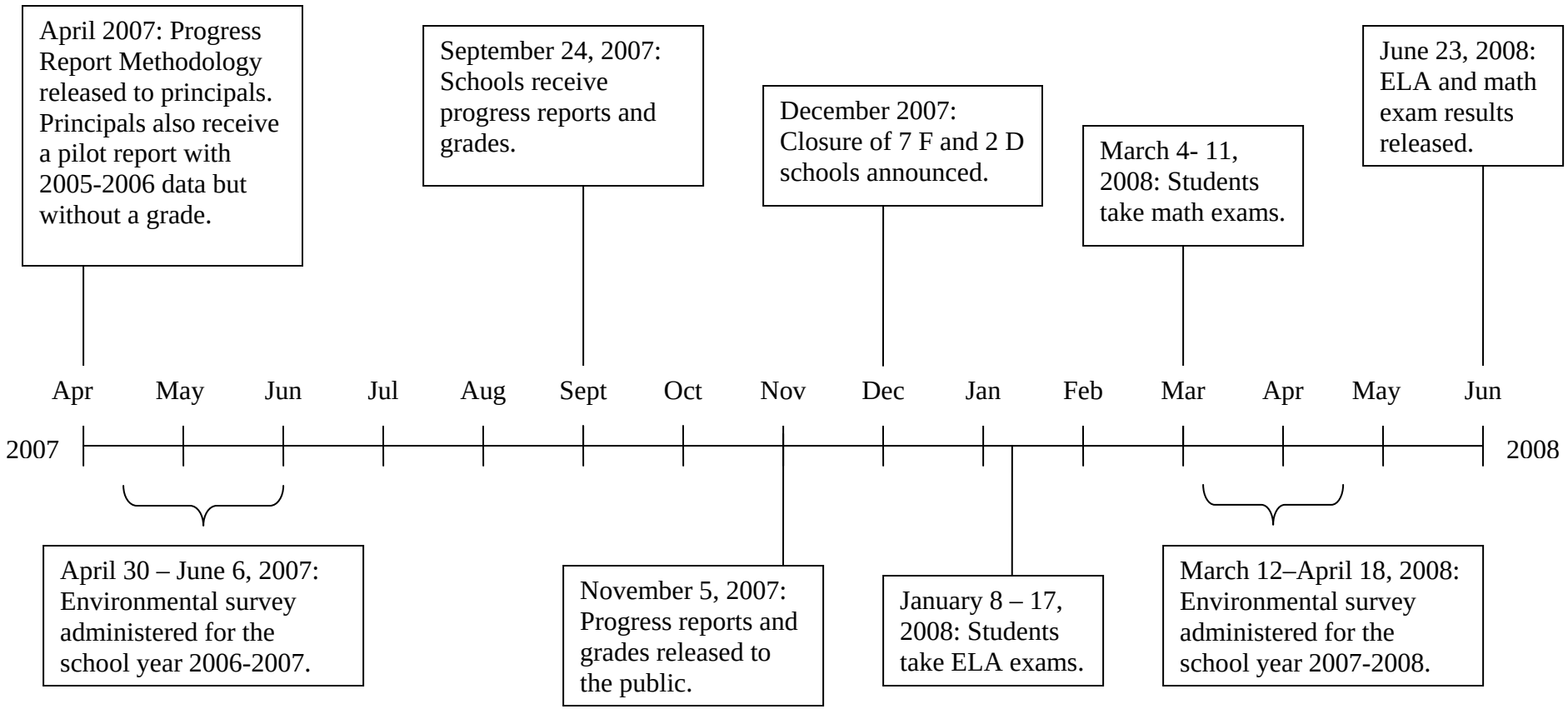
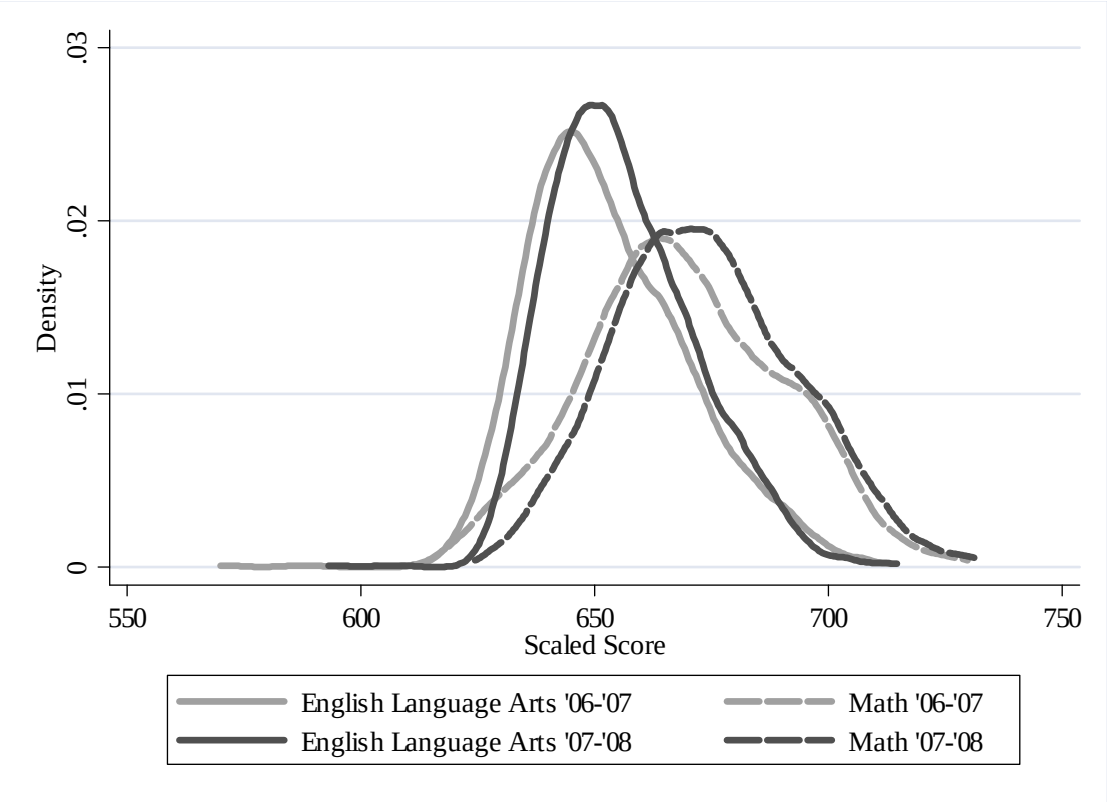
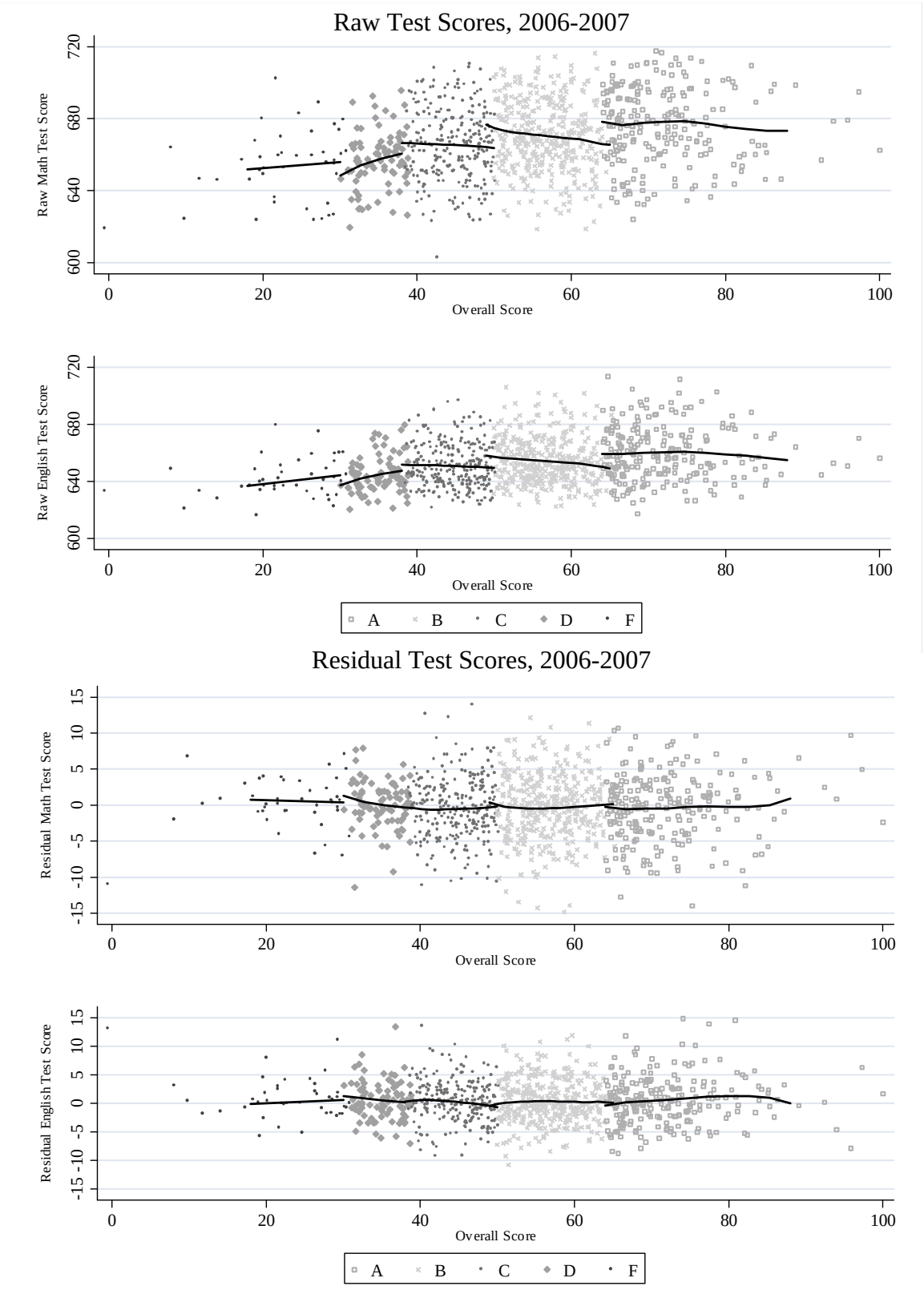


Figure 3: Distribution of School Average Test Scores in 2006-2007 and 2007-2008



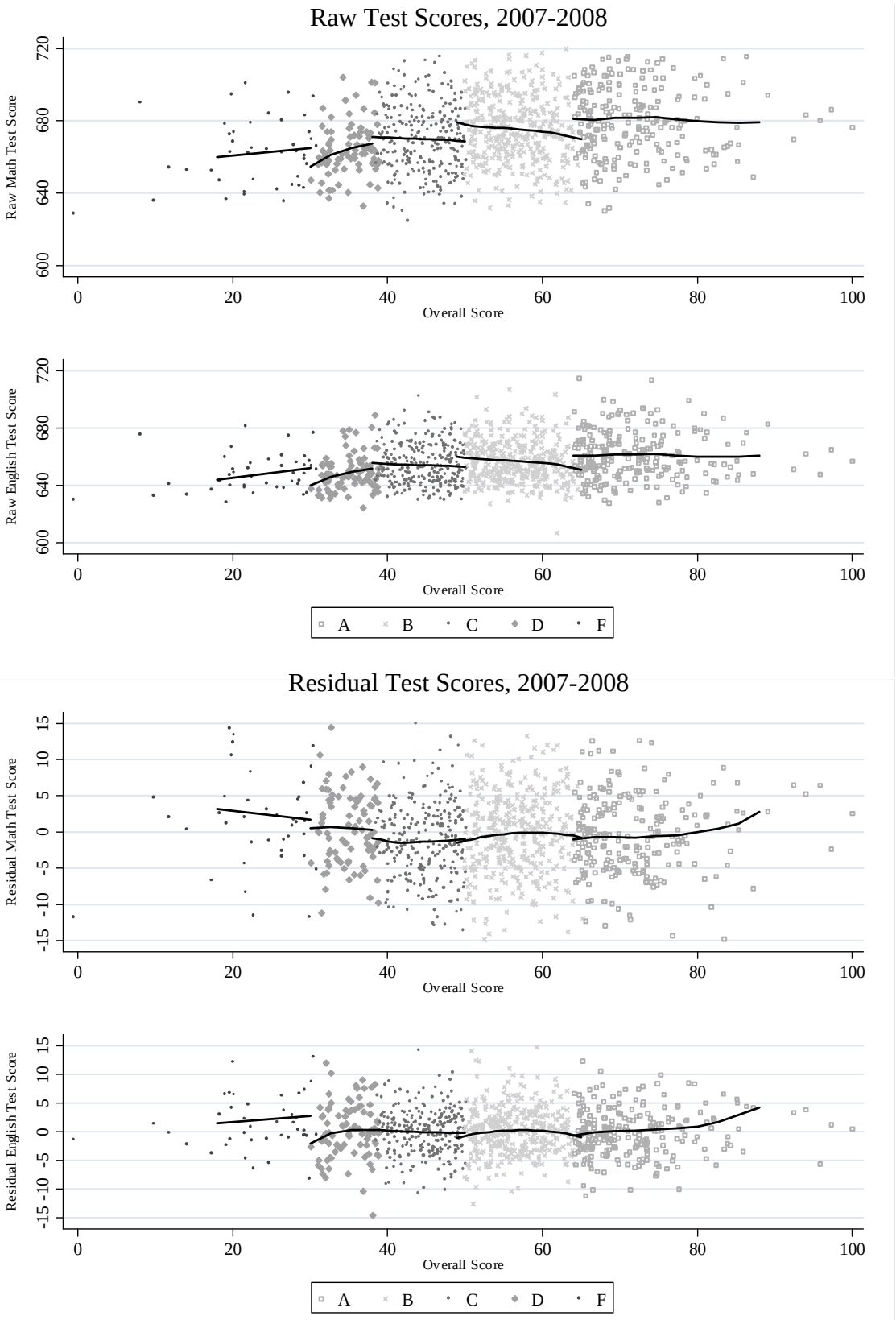
Note: Densities shown for the 985 schools used in our empirical analysis.

Figure 4: School Average Math and English Scale Scores by Progress Report Grade



Note: Residuals from a regression of test scores on the components of the overall score: peer index, progress, performance, environment, and additional credit. We also include quartic polynomial for each component, and the impact of these variables is allowed to vary across the three types of schools (Elementary, K-8, and Middle). The solid lines plot estimates from a locally weighted polynomial (Fan) regression with a bandwidth of 8 points performed separately within each group of schools with the same progress report grade.

Figure 5: School Average Math and English Scale Scores by Progress Report Grade



Note: Residuals from a regression of test scores on the components of the overall score: peer index, progress, performance, environment, and additional credit. We also include a quartic polynomial for each component, and the impact of these variables is allowed to vary across the three types of schools (Elementary, K-8, and Middle). The solid lines plot estimates from a locally weighted polynomial (Fan) regression with a bandwidth of 8 points performed separately within each group of schools with the same progress report grade.

Table A1: Survey Questions used by Respondent Group and Year

	Parents		Teachers		Students	
	2007	2008	2007	2008	2007	2008
Course Offerings	8 & 9	6a-f, 7a-f	7a-f	3a-f	9a-f, 10a-f	9a-f, 10a-f
High Expectations	7a	5a	6a-b	2a-b	3b-d, 3g	3b-d, 3g
Teacher Quality	14a	13a	--	--	--	--
Overall Satisfaction with Education	14d	13e	--	--	--	--
Focus on Teaching by School Leaders	--	--	10f-h	6g-i	--	--
Use of Student Data	--	--	6f & 11b	6k & 7a	--	--
Professional Development Quality	--	--	11c-d	7b-c	--	--
Quality of Instructional Materials	--	--	6d-e	7d-e	--	--
Essays and Projects	--	--	--	--	7a-b	7a-b
Group and Hands-on Learning Activities	--	--	--	--	8b-d	8b-d