Discipline Reform, School Culture, and Student Achievement

Ashley C. Craig
Australian National University

David C. Martin
Harvard University

March 2025

Abstract

Does relaxing strict school discipline improve student achievement, or lead to classroom disorder? We study a 2012 reform in New York City public middle schools that eliminated suspensions for non-violent, disorderly behavior. Math scores of students in more-affected schools rose by 0.05 standard deviations over three years relative to other schools. Reading scores rose by 0.03 standard deviations. Only a small portion of these aggregate benefits can be explained by the direct impact of eliminating suspensions on students who would have been suspended under the old policy. Instead, test score gains are associated with improvements in school culture, as measured by the quality of student-teacher relationships and perceptions of safety at school. We find no evidence of trade-offs between students, with students benefiting even if they were unlikely to be suspended themselves.

JEL codes: H75, I2, J24, J45

Keywords: education, school suspension, school discipline, school safety, human capital

---

Ashley Craig: ashley.craig@anu.edu.au; Research School of Economics, R2094, LF Crisp Building, 25a Kingsley St, Acton ACT 2601

David Martin: david_martin@fas.harvard.edu

Declarations of interest: None.

Thanks: We thank Brian Hall, Lawrence Katz, Amanda Pallais, and Andrei Shleifer for their comments, guidance, and support. We are also indebted to Isaiah Andrews, Andrew Bacher-Hicks, Iwan Barankay, Alex Bell, Sophie Calder-Wang, Raj Chetty, Olivia Chi, David Deming, Roland Fryer, Edward Glaeser, Joshua Goodman, Nora Gordon, Nathaniel Hendren, Sara Heller, James Hines, Peter Hull, Kosuke Imai, Michael Mueller-Smith, Jonathan Roth, Jeffrey Smith, Rucha Vankudre, and seminar participants at Harvard University, the University of Michigan, Pomona College, Western University, the University of Rochester, UNSW, the University of Queensland, the University of Melbourne, the E61 Institute, the Australian National University, and the 20th IZA/SOLE Transatlantic Meeting. This project would not have been possible without the data and wealth of knowledge we received from dedicated public servants at the New York City Department of Education, especially Mabel Fu, Derek Li, Sophie Sharps, and Joshua Smith. Johanna Miller and numerous unnamed educators provided valuable institutional knowledge. Financial support from an Inequality and Wealth Concentration Ph.D. Scholarship from the Program in Inequality & Social Policy at Harvard University is gratefully acknowledged. This research was approved by institutional review boards at Harvard University (# IRB16-2124) and the University of Michigan (# HUM00163349).

Success should not be measured by the number of suspensions, but by the number of schools with an improved school climate.

Michael Mulgrew
President of the United Federation of Teachers

1 Introduction

School discipline reforms have increasingly sought to limit the use of suspensions, which are punishments that exclude misbehaving students from class. These reforms are controversial. High suspension rates have been linked to low test scores, unsafe schools, high drop-out rates, and increased criminal activity (Steinberg *et al.*, 2011; Perry and Morris, 2014; Noltemeyer *et al.*, 2015; Bacher-Hicks *et al.*, 2019).[1] But suspensions may help insulate other students from negative consequences of disruptive classroom behavior (Carrell and Hoekstra, 2010; Lavy and Schlosser, 2011; Carrell *et al.*, 2018). The threat of harsh punishment may also deter misbehavior in the first place. Indeed, strict discipline has been suggested as a reason why high-performing "no excuses" charter schools are so effective in raising test scores (Angrist *et al.*, 2013). Some have therefore argued that banning suspensions could actually harm both school culture and academic achievement (Disare, 2016; Gregory *et al.*, 2010; Zimmerman, 2016; Steinberg and Lacoe, 2017). Finding the right balance has important implications for human capital development.

We study a 2012 reform in New York City public middle schools that eliminated suspensions for non-violent, disorderly behavior.[2] Schools were forced to replace these suspensions with less disruptive punishments such as removal from a single class, and were encouraged to employ non-punitive interventions. For two groups of students, our analysis tracks test scores, behavior, and measures of school culture such as assessments of student-teacher relationships. The *High Treatment* group includes middle school students in school-grade cells with above-median historical suspension rates for disorderly behavior, who were thus more affected by the reform. The *Low Treatment* group contains those in school-grades with below-median pre-reform suspension rates for these types of behavior, where such suspensions had rarely or never been used. Aside from suspension rates, these groups are similar in observable characteristics.

Using administrative data on students in grades 6-8 from the New York City Department of Education (NYCDOE), we adopt a *difference-in-differences* framework to exploit the sharp timing of the 2012 discipline reform. Over the four years prior to the reform, average test scores and suspension rates for disorderly behavior moved largely in parallel

[1] Moreover, suspension rates for black and Hispanic students are much higher than for white students. To the extent that high suspension rates are causally harmful, disproportionate discipline may contribute to differences in achievement between white and minority students (Morris and Perry, 2016; Welsh, 2023).

[2] Examples of non-violent, disorderly behavior include profane language and persistent non-compliance. It is an open question whether our results generalize to suspensions for fighting or other violent behavior.

in High and Low Treatment school-grades. Suspension rates for disorderly behavior then fell sharply to zero in both groups in 2012, but the drop was three times as large in the High Treatment group as in the Low Treatment group. Our key identifying assumption is that average test scores in the two groups would have continued to move together if the reform had not been implemented.

Over the three years following the reform, average math scores for students in the High Treatment group rose by 0.05 standard deviations relative to the Low Treatment group. Reading scores rose by 0.03 standard deviations. These improvements are large, given that the reform came at minimal financial cost. For comparison, the test score gains are equivalent to raising teacher quality by one third of a standard deviation (Chetty *et al.*, 2014). As we describe below, the benefits were shared by a broad range of students.

Our results are robust to many alternative specifications, including different treatment definitions, allowing for linear differences in pre-trends, and balancing treatment groups on demographics. We also rule out the possibility that they are driven by two major policy changes around the time of the reform. Specifically, the timing of the switch to Common Core testing did not coincide with the test score gains we see, and the impact of the reform preceded the election of Mayor Bill de Blasio and his appointment of Chancellor Carmen Fariña. Finally, we test several specific channels through which those or any other policy changes could have operated, including differential replacement of principals or relative changes in funding between the two treatment groups.

The achievement gains we see combine *direct effects*, *behavioral effects*, and *spillovers*. Direct effects arise from replacing suspensions with alternative interventions, holding student behavior fixed. This could lead to increased instructional time for those who would have been suspended, or the elimination of the stigma and psychological costs of suspension. Behavioral effects arise when students change their own behavior in response to the disciplinary regime. They could go in either direction: Reducing the threat of suspension could induce more misbehavior, but de-emphasizing punishment could—as we ultimately find—improve school culture and thus reduce misbehavior.[3] Finally, both punishments and changes in behavior have the potential to cause spillovers between students. For example, keeping disruptive students in the classroom could prevent other students from learning. But positive changes in students' behavior could benefit their peers.

We find little variation in the effect of the reform between students who are more or less likely to be suspended, which suggests that the gains are not driven by direct effects from

[3]There are many reasons why a harsher discipline code could produce worse behavior, which are hard to separate. Punishment could undermine intrinsic motivation to behave well (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2003, 2006), change students' perceptions of themselves in harmful ways (Ellingsen and Johannesson, 2008), or—as suggested by our results below—undermine student-teacher relationships.

replacing suspensions with other punishments. When we use a rich set of demographics and baseline test scores to predict each student's risk of suspension, treatment effects are similar across students with very different suspension likelihoods. Reflecting this, test score gains for boys and girls are nearly identical, despite boys being suspended twice as often as girls; and gains are actually smaller for black students than for white or Hispanic students, despite black students being suspended far more frequently.

The test score gains are better-explained by improvements in school culture induced by the 2012 reform. Student and teacher survey responses reveal that the quality of student-teacher relationships and perceptions of safety both improved in High Treatment school-grades relative to Low Treatment school-grades. School-grades with larger improvements in culture achieved systematically larger test score gains. Moreover, back-of-the-envelope calculations based on the cross-sectional relationships between our measures of culture and test scores suggest that the effects on culture are large enough to explain the entire impact on achievement. These results help explain why we see achievement gains for a broad range of students, rather than only those who would have been most likely to be suspended. They are also consistent with data on reported incidents of disruptive behavior, which suggest that there were measurable behavioral improvements in High Treatment school-grades to match student and teacher perceptions.

This pattern of results is consistent with our conversations with teachers and reform advocates, who highlighted several mechanisms through which the relaxation of school discipline could affect school culture, and flow through to student achievement. First, both students and teachers may work harder when they feel supported and respected. Second, students and teachers may reallocate effort away from avoiding and enforcing discipline. Third, students may become more engaged if they perceive that teachers and administrators are less biased and more reasonable.

By contrast, directs effects are too small on a per-suspension basis to explain our aggregate treatment effects, given that most students are never suspended. We use the sharp timing of each suspension relative to standardized exams to show that suspending a student for disorderly behavior has at most a 0.03 standard deviation negative effect on their math score that year. By contrast, if the achievement gains from the reform came solely from the elimination of the direct effect of each suspension, the impact per suspension must have been over 8 standard deviations. Such a large causal impact is implausible in light of our estimates. In fact, this implied effect per suspension is far larger even than the cross-sectional relationship between suspension and test score performance.

Our results address a key aspect of the policy debate surrounding school discipline. Much of the disagreement between proponents and opponents of discipline reform can

be traced to beliefs about the relationship between strict, punitive discipline and school culture. Indeed, a stated justification for zero tolerance school discipline was that it was necessary to maintain a safe school environment that is conducive to learning (Skiba and Knesting, 2001; American Psychological Association Zero Tolerance Task Force, 2008). Similarly, some have argued that reforms limiting suspension use have reduced safety and increased disruption by making it difficult for teachers to manage misbehavior (Federal Commission on School Safety, 2018; Eden, 2017). This idea is supported in principle by evidence of large negative spillover effects from disruptive students on their peers (Carrell and Hoekstra, 2010; Carrell *et al.*, 2018). Similar logic underlies the strict discipline policies of No Excuses charter schools, the practices of which have been shown to jointly raise student achievement (Angrist *et al.*, 2013; Dobbie and Fryer, 2013).

The potential for improvements in school culture has also underpinned arguments for *relaxing* school discipline. This was an important basis for the push by the New York City Civil Liberties Union (NYCLU) and other advocacy groups that led to the 2012 discipline reform we study in this paper (Mukherjee, 2007; Miller *et al.*, 2011). On a national scale, the Departments of Justice and Education issued a *Dear Colleague* letter to schools in 2014, urging them to reduce suspension use. They argued that doing so would improve school culture by reducing perceptions that discipline policy was biased and unduly harsh. It has been difficult for researchers to prove that relaxing discipline has a positive causal impact on school culture, but our results provide support for this view.[4] We find that relaxing discipline improved safety, student-teacher relationships, and test scores. The policy therefore benefited a wide range of students, with no evidence of trade-offs between students with different characteristics or likelihoods of being punished with a suspension.

School Discipline in the Literature

This paper complements work by Bacher-Hicks *et al.* (2019) in Charlotte-Mecklenberg schools, which harnesses quasi-random assignment of students to schools with different suspension rates. They find that being assigned to a school with a higher suspension rate has negative effects on long-run outcomes such as graduation and criminal activity. One advantage of our approach is that we study a reform to discipline policy specifically, rather than measuring the effect of going to a school with more suspensions. Put differently, we circumvent the main limitation of Bacher-Hicks *et al.*'s analysis, which is that differences in suspension rates across schools may be correlated with other factors that also affect stu-

[4]Teachers and students feel less safe in schools with high suspension rates, but a causal relationship is harder to establish (Skiba and Knesting, 2001; Lacoe and Steinberg, 2018). Similarly, student achievement and school culture have been linked, but there is only limited evidence showing causality (Brookover *et al.*, 1978; Pallas, 1988; Kutsyuruba *et al.*, 2015; Kraft *et al.*, 2016; Dulay and Karadağ, 2017; Pas *et al.*, 2019). Finally, Backes *et al.* (2022) measure "climate value-added" and show that it is related to test score value-added.

dent outcomes. Instead, we rely on an assumption that there are no unobserved shocks that differentially affect schools with higher or lower suspension rates following the 2012 discipline reform.[5] We probe this assumption in detail below.

We also build on evidence from Philadelphia, where Lacoe and Steinberg (2019) use a similar discipline code change as an instrument for suspension.[6] They find that suspensions negatively affect the test scores of both suspended students and their peers, under the assumption that students who had been suspended in the past would have been suspended again in the absence of the reform. Unfortunately, with only two years of data, they are limited in their ability to assess the validity of their empirical design. Related work by Lacoe and Steinberg (2018) suggests that the same reform increased truancy rates, despite having little impact on the total suspension rate in Philadelphia schools.[7] Building on our study of New York City, Cleveland (2022) finds gains in reading scores for some students from a statewide reform in Massachusetts. By contrast, Pope and Zuo (2023) find that a gradual fall in suspension rates over a decade in Los Angeles was associated with lower test scores in the schools most affected. This difference in results may stem from the gradual and less-centralized nature of the L.A. reform.

Kinsler (2013) takes a structural approach. Based on a model that accounts for potential spillovers from disruptive behavior, his calibration suggests that stricter discipline may have a positive effect on student performance through improvements in behavior. A critical assumption underlying this model is that principals set policy optimally, weighing potential harms to students who are suspended against benefits to others from improved behavior. Our reduced-form approach avoids this assumption. Our results suggest that behavior improves rather than deteriorates as discipline policy is relaxed, with no evidence of trade-offs between different types of students.

More broadly, we contribute to a literature spanning economics, sociology and education policy studying the associations between suspensions and test scores, drop-out rates and criminal behavior (Gregory *et al.*, 2010; Hinze-Pifer and Sartain, 2018; Anderson *et al.*, 2019; Sorensen *et al.*, 2022).[8] For example, Cobb-Clark *et al.* (2015) argue that the relationship between suspensions and achievement can be fully explained by individual characteristics if the degree of selection on observables is similar to selection on unobservables.

[5] This assumption needs to be strengthened if treatment effects are heterogeneous (see Section 4).

[6] Several other studies examine the effects of suspension reforms on the rate and composition of suspensions (Baker-Smith, 2018; Craigie, 2022; Hashim *et al.*, 2022).

[7] Small impacts on the suspension rates can be due to compliance problems (Anderson, 2018; Anderson and McKenzie, 2022). We find a meaningful impact on the suspension rate as intended. This is in line with Baker-Smith (2018), who focuses on the impact of the same reform on the composition of suspensions.

[8] Most of this work is empirical, but there are theoretical contributions (Lazear, 2001). In turn, this is part of a vast literature studying education production functions more generally (Fryer, 2017; Hanushek, 2020).

However, Perry and Morris (2014) suggest that suspensions adversely affect the peers of suspended students, based on regressions that control for student and school fixed effects. Noltemeyer *et al.* (2015) provide a broad meta-analysis of previous studies.

The paper proceeds as follows. We describe the institutional setting and the school discipline reform we use for identification in Section 2. This is followed by a summary of our data, sample restrictions, and descriptive statistics in Section 3. Our analysis then proceeds with our estimation of the effects of that reform on achievement in Section 4. In Section 5, we present evidence that these achievement gains are explained by improvements in school culture which benefits a broad range of students, including those who would not likely have been suspended. We dedicate Section 6 to ruling out other specific mechanisms, including any direct effects of missed instructional time, principal and teacher turnover, and changes in resources. Section 7 concludes.

2 School Discipline in New York City

Discipline in New York City public schools is governed by the Citywide Standards of Intervention and Discipline Measures ("the discipline code"). There are five levels of disciplinary infractions.[9] Level 1 infractions cover non-compliance such as being late for class, excessive noise, and disrespectful behavior. At the other extreme, Level 5 infractions are for severe misbehavior, which is usually dangerous or violent. The reform we study targeted Level 2 infractions. These are for disorderly behavior such as profane language and persistent non-compliance. Although these levels are associated with *infractions*, we occasionally refer to a suspension for Level x behavior as a "Level x suspension."

For infractions at each level, the discipline code prescribes a range of allowable interventions. As a first step, schools may use non-punitive guidance interventions such as counseling, which encourage positive behavior rather than punishing negative behavior. If these are insufficient, punitive disciplinary measures may be taken. These range in severity from admonishment by a teacher to, in very limited cases, expulsion.

Three common punishments result in temporary exclusion from classroom instruction. First, students may be removed from a single class, which still allows them to attend other classes that day. Second, students may receive a *principal's suspension*, in which case they miss all classes for between one and five days. Third, for Level 3 infractions and above, schools can escalate the matter to the district-wide Office of Safety and Youth Development (OSYD) to request a *superintendent's suspension*; these last for 6 days or longer. During classroom removals and principal's suspensions, students receive alternative instruc-

[9] Details about each infraction level, suspension length, and reasons are available in Appendix Table I1.

tion at a different location within their school. Superintendent's suspensions are served at an alternative learning center or buddy school (New York City Department of Education, 2004). There are therefore no true "out-of-school" suspensions in New York City schools. However, the quality of the alternative instruction students receive is generally perceived to be an imperfect substitute for normal classroom instruction.[10]

In practice, implementation of the discipline code varies with the preferences of teachers and school administrators. Teachers are generally the first decision-makers when students misbehave, choosing whether to handle situations internally or report them to school administrators. More broadly, teachers set the tone for what is acceptable behavior through thousands of day-to-day interactions with their students. In doing so, teachers follow the lead of their principals, who are ultimately responsible for student discipline.[11] Some principals choose to interpret district guidelines more strictly than others. Appendix Figure I3 provides a simplified flow chart of the process that leads to a suspension.

Overall, discipline policy in New York City schools grew stricter over the late 1990s and 2000s. In 1998, the city's Board of Education transferred responsibility for school safety to the New York Police Department, echoing then-Mayor Rudolph Giuliani's "tough on crime" message. Thus began a period of aggressive discipline policy that was continued by Giuliani's successor, Michael Bloomberg. In 2004, Bloomberg enacted the Impact Schools Initiative, which doubled police presence at targeted high-crime schools and called for zero tolerance toward disciplinary infractions. In 2006, he introduced a roving metal detector program at all middle and high schools. Over this period, the number of suspensions per year increased rapidly, peaking at nearly 74,000 in 2008 (see Figure 1).

In the 2010s, the NYCDOE began easing its reliance on suspensions in response to lobbying efforts from groups such as the New York Civil Liberties Union (NYCLU) and Teachers United, who argued that existing policies disproportionately harmed minority students.[12] We study a reform in 2012 that was the first major step in this process. Specifically, the 2012 reform prohibited suspensions for disorderly behavior (Level 2 infractions).[13] Figures 1 and 2 show the rates of suspension overall and by infraction level for each school year. Here and throughout the paper, we use 2012 to refer to the 2012-13 school year, and so forth. These two figures show that the prohibition had an immediate

[10]District regulations require alternative instruction to include class and homework assignments, and to provide students the opportunity to continue earning academic credit (New York City Department of Education, 2004). But our conversations with teachers and administrators suggest that coordination between primary and alternative educators is often poor, and that attendance at alternative learning centers is low.

[11]Teachers and administrators with whom we have spoken emphasize that principals shape the discipline cultures of schools. However, some principals delegate decisions regarding individual suspensions.

[12]Our data confirm that the suspension rate is higher for minority students in New York City. Between 2006 and 2016, black students made up 30 percent of enrollment but 45 percent of suspended students.

[13]Suspensions had already been prohibited for Level 1 infractions.

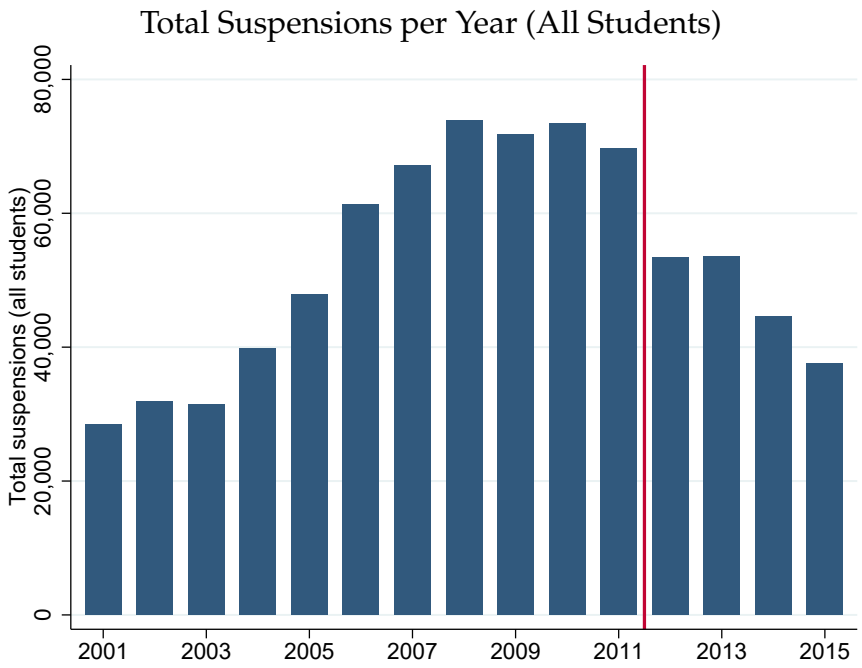

Figure 1. This figure shows the total number of principal's and superintendent's suspensions issued per year in the New York City public school system, including all suspensions and all students. The vertical red line indicates the timing of the reform. 2012 refers to the 2012-13 school year, and so forth. Data come from reports by the New York Civil Liberties Union.

impact: Suspensions for Level 2 infractions fell to zero except for extremely rare cases, and the total suspension rate fell sharply.

Comparison to Other States

Although Figure 1 shows that suspension rates were historically high in New York City before this reform, many other states had higher suspension rates than New York. Figure I4 displays civil rights data collected by the United States Department of Education. Bearing in mind that these state-level data include non-urban schools and cover all grades, we can see that New York State is roughly comparable to Massachusetts and California, but more than twice as many students receive a suspension in Florida and Mississippi.

3 Data on School Discipline, Culture, and Achievement

To study the link between discipline and achievement in New York City, we combine administrative data on students and staff with public data on school culture, violent and disruptive incidents, and funding.

3.1 Administrative Data

Our main data source is comprised of administrative records from the NYCDOE covering all students in New York City public schools. New York City is the largest public school

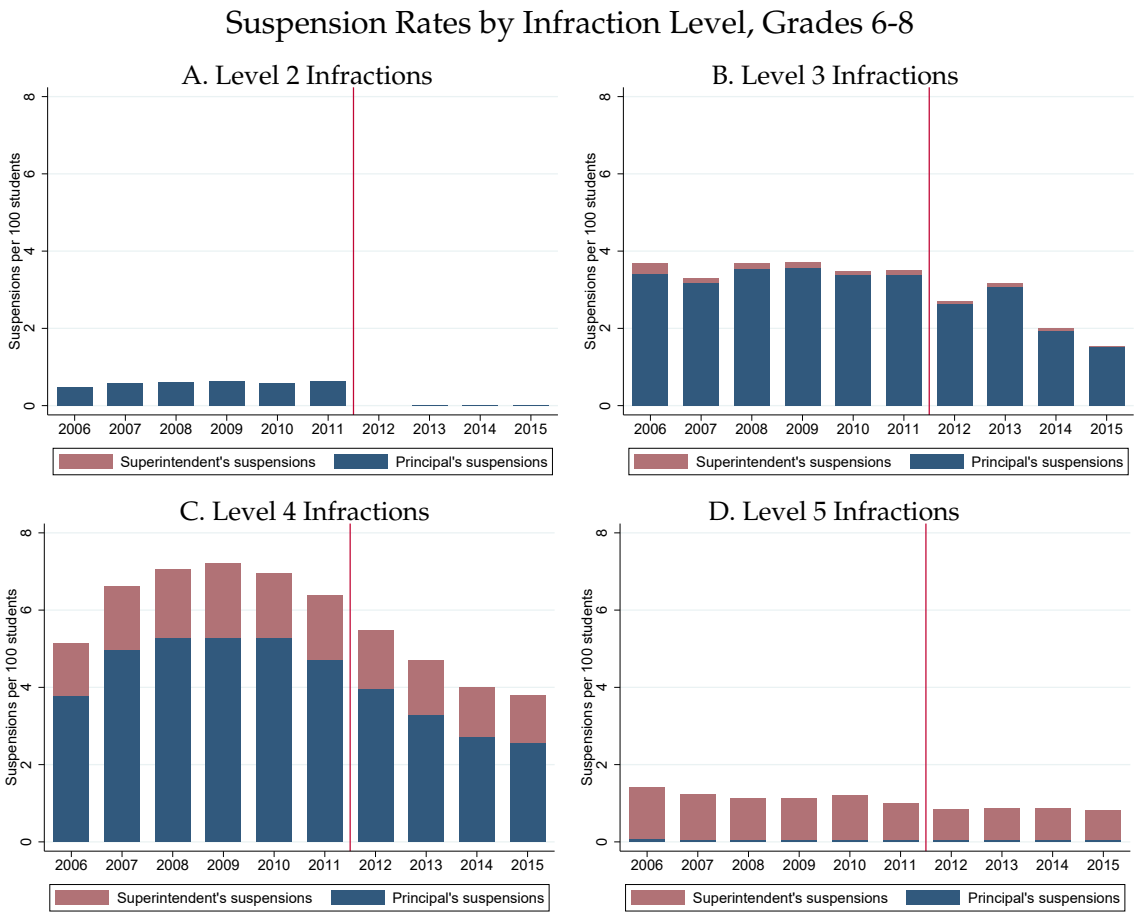

Figure 2. This figure shows yearly suspension rates by infraction level for students in grades 6-8 in New York City public schools. Suspension rates are expressed as the number of suspensions per 100 students per year. Each panel contains suspensions for a given infraction level, with principal's suspensions (1-5 days) in blue and superintendent's suspensions (6+ days) in red. The vertical red lines indicate the timing of the reform. 2012 refers to the 2012-13 school year etc. Data are from the New York City Department of Education.

district in the United States, currently responsible for educating over 1.1 million students in over 1,800 schools. From 2006 to 2015, we observe student demographic information, enrollment, attendance, disciplinary events, annual test scores for grades 3 through 8, and subject-specific test scores for higher grades.[14] We also observe yearly staffing rosters with information on teachers and other school employees.

Discipline

We have data on suspensions from the NYCDOE's Suspensions and Office of Hearings Online system. Records include the start and end dates of each suspension, an infraction code, and scrambled identifiers to link suspensions to other NYCDOE data. However, we do not have data on other disciplinary responses such as classroom removals and guid-

[14]Throughout, we refer to the school year beginning in September 2006 as 2006, rather than 2006-07. From 2000 to 2005, we observe annual test score records for grades 3-8, but none of the other variables. We use these earlier test scores to link older students in our sample to their test scores from previous grades.

ance interventions. We also see yearly infraction counts by student from the NYCDOE's Online Occurence Reporting System: for each student, we observe the total number of Level 1 through Level 3 infractions, and the total of Level 4 and Level 5 infractions.

Test scores

Each year, students in grades 3-8 take statewide exams in mathematics and "English / Language Arts" (ELA), which we refer to as "reading." We standardize reported test scores to have zero mean and unit standard deviation within each grade-subject-year cell, which facilitates comparisons across tests. In Appendix E.6, we repeat our primary analysis using test score percentiles, with very similar results.

In 2012, state assessments were changed to align with a nationwide set of academic standards referred to as the *Common Core*. The timing of this change coincides with the reform that we study. However, the switch to the new tests is unlikely to have affected our results. Although a new testing methodology could in principle yield better or worse assessments of students' knowledge, any differential impact on our two treatment groups would have been felt as a sharp change in 2012 rather than the gradual impact that we observe. We discuss this issue further in Appendix F.

Students, teachers, and other school personnel

We observe student race and gender. We also see an indicator for whether the student has learned English as a second language (ELL), or has an Individual Education Plan due to special needs. As is standard in the education literature, we use an indicator of eligibility for free or reduced price lunch as a proxy for family income.

Data on teachers and other school personnel come from the NYCDOE's human resources system. Records include salary, tenure at a given school, and total years in the district. We can tabulate counts of teachers, counselors, and psychologists. These variables help us better understand the mechanisms that drive our aggregate results.

Sample

We focus on middle school students in grades 6 to 8. Although annual test scores are available for students in grades 3 to 5, these students are subject to a different disciplinary regime, and were affected by a contemporaneous policy change in 2012 that reduced the maximum length of suspensions for Level 3 behavior from 10 to 5 days. Because we have no credible way to delineate groups who were more or less affected by this change, we focus on students in grades 6 to 8, who were unaffected by this parallel reform.

We further restrict to school-grade combinations that appear in our data in every year. This yields a balanced panel at the school-grade level, which is the level of treatment in our quasi-experimental analysis. This balance ensures that selection of school-grades into

and out of the sample does not affect our results. There is, however, no way to balance our sample at the individual student level, since we observe each student for a maximum of three years (grades 6 to 8) out of a ten-year period. We discuss balance on student characteristics in Section 4.1 and Appendix E.1.

Over our sample period, enrollment in charter schools increased from 1.3% to 9.7% of New York City students in grades 6-8. However, the NYCDOE does not collect disciplinary records for charter schools because they set their own discipline policies. We therefore omit charter school students from the analysis. In addition, we omit all home-schooled students and students in special-education-only schools. We show in Appendix G that our results are not driven by selective sorting of students.

Throughout the paper, we include all students who satisfy these restrictions. However, in Appendix E.8, we exclude accelerated students who go on to complete the New York State Regents exam in math one year early (in grade 8). We do this to rule out any possibility that our results are driven by a testing waiver that was received by the NYCDOE from the United States Department of Education in 2013 to avoid 'double testing' accelerated students via both New York State Regents exams and the standardized exams normally administered in grades 6 through 8. Our results are unaffected.

3.2 School Characteristics

We supplement our administrative data with student and teacher responses to the NYCDOE's annual survey on the school environment, reports from the New York State Violent or Disruptive Incident Reporting (VADIR) system, and data on school funding.

Survey Responses

Students in grades 6-12, their teachers, and their parents complete an annual survey designed to assess schools' efficacy in supporting student success. We use answers to these surveys to measure important dimensions of school culture: the quality of student-teacher relationships, student behavior, and feelings of safety at school. We provide a detailed explanation of how we use the survey data in Section 5, and present evidence that improvements in culture are important in driving the gains from the 2012 reform.

Violent or Disruptive Incidents

Federal law requires that all New York City schools submit annual reports of violent or disruptive incidents to the state through the VADIR system. These reports are used to calculate a School Violence Index (SVI) that weights incidents by their severity and scales by enrollment, which is used to classify schools as "persistently dangerous." Schools also report "other disruptive incidents," which do not enter into the SVI calculations.

Funding

We obtained public, school-level data on per-student school expenditures compiled by the Research Alliance for New York City Schools, an organization based out of New York University's Steinhardt School of Culture, Education, and Human Development. These data include total expenditures, as well as expenditures broken down into classroom instruction, instructional support services, leadership/supervision/support, ancillary support services (food, transportation, school safety, building services), and other costs.

3.3 Descriptive Statistics

Table 1 shows summary statistics for our sample of middle school students for the period prior to the discipline reform of 2012. As a share of all enrollment, 15 percent of students are white, 28 percent are black and 40 percent are Hispanic. Although 68 percent of all middle school students are eligible for free or reduced-price lunch, poverty rates are much higher among black and Hispanic students than white students. We also see evidence of academic achievement gaps by race and by sex: the unconditional gaps between average test scores of white and black students are 0.63σ in reading and 0.72σ in math; girls score 0.25σ better than boys in reading and 0.08σ better in math.

Across all students and infractions, there were over 11 suspensions per 100 students per year from 2006 to 2011.[15] Over half of these are for Level 4 infractions, for aggressive or injurious behavior. Suspensions for Level 2 infractions, which provide us with our quasi-experimental variation, make up five percent of the total. Like test scores, suspension rates vary dramatically by sex and race. Boys are suspended more than twice as often as girls (16 per 100 students, compared to 7); and black students are suspended nearly 2.5 times as often as white students (18 per 100 students, compared to 7).

Black and Hispanic students attend schools with more resources than white students, as measured by traditional school inputs. Their schools spend almost $2,000 more per student and have smaller class sizes, although their teachers earn lower salaries. This distribution of resources may reflect efforts to address racial achievement gaps, either directly or by targeting low-income or low-SES students.

Relationship between suspensions and test scores

Figure 3 shows the powerful negative unconditional relationship between suspensions and test scores in our data. For both math and reading, suspension rates fall as exam performance rises: students with a one percentile higher math score are 0.3 percent less likely to be suspended, with only a slightly weaker association between suspension and reading

[15] About two-thirds of all suspensions go to students who only get suspended once per year. See Appendix Figure I2 for a breakdown of annual suspension frequencies for suspended students.

Table 1. Summary Statistics for Grades 6-8, 2006-2011

	All Students	Boys	Girls	Black	Hispanic	White	Other
Share of Enrollment	1.000	0.509	0.491	0.280	0.404	0.150	0.166
% Free Lunch	0.676	0.674	0.679	0.724	0.778	0.376	0.621
% ELL	0.127	0.139	0.114	0.027	0.208	0.058	0.159
Test Scores (Math)	-0.000	-0.038	0.040	-0.298	-0.237	0.426	0.683
Test Scores (Reading)	-0.000	-0.122	0.125	-0.154	-0.232	0.471	0.397
Susp. / 100 (All)	11.3	15.6	6.8	17.8	11.3	7.3	3.5
Susp. / 100 (Level 2)	0.6	0.8	0.3	0.9	0.6	0.4	0.1
Susp. / 100 (Level 3)	3.5	4.8	2.1	5.3	3.5	2.6	1.1
Susp. / 100 (Level 4)	6.1	8.4	3.7	9.7	6.2	3.9	2.0
Susp. / 100 (Level 5)	1.1	1.5	0.6	1.9	1.0	0.4	0.3
Expenditure / Student	16,885	16,927	16,842	17,445	17,403	15,790	15,671
Teacher Salary	68,707	68,757	68,655	68,320	67,513	70,921	70,268
Avg. School-Grade Size	292.5	295.0	289.9	236.1	281.0	339.2	373.2
Students per Teacher	14.3	14.3	14.4	13.9	13.8	15.5	15.4
Students per Counselor	286.9	284.6	289.3	275.8	258.1	335.6	331.2

Table notes. This table shows summary statistics for the pre-reform period (2006-2011) in our analysis sample, which includes students in grades 6-8 who do not attend charter schools, schools that opened or closed between 2006 and 2015, or special-education-only schools. Test scores are standardized within this sample in subject-grade-year cells. Suspension rates are expressed as the number of suspensions per 100 students. Expenditure data are from the Research Alliance for New York City Schools; other data are from the New York City Department of Education.

scores. This relationship is consistent with suspensions being causally harmful. However, it is also consistent with students who are disadvantaged or less able performing worse on standardized tests for those reasons, while also behaving badly.

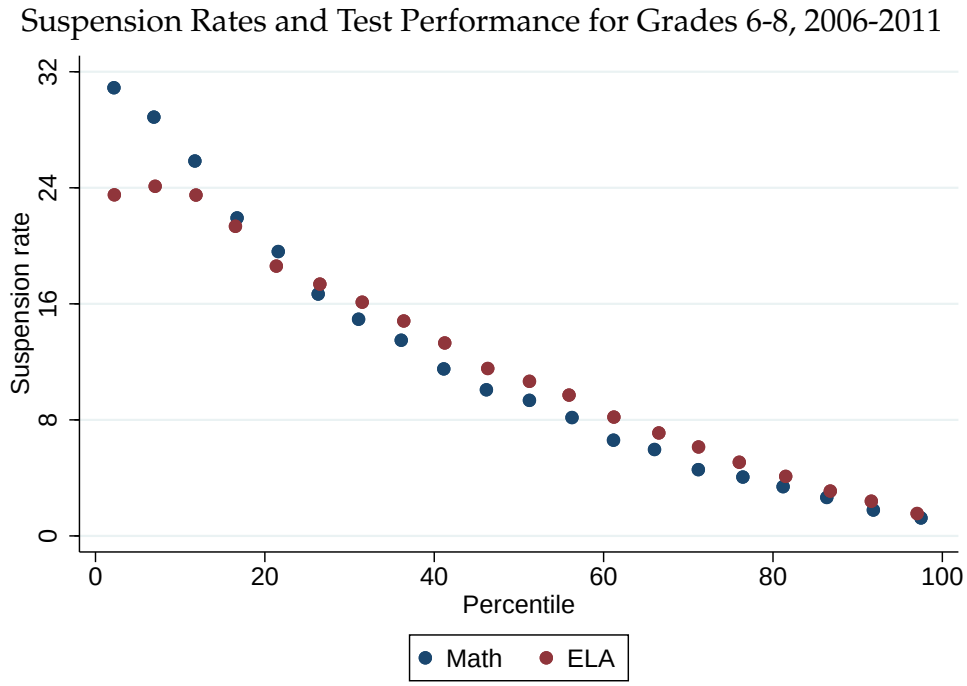

Figure 3. This figure shows the relationships between suspension rates and contemporaneous test scores in math and reading (ELA) for our sample of middle school students (grades 6-8) prior to the discipline reform. Students are grouped into bins based on their percentile scores. The figure shows the number of suspensions per 100 students in each bin plotted against the average percentile score in each bin. Data are from the New York City Department of Education.

In Appendix A, we use regressions with fixed effects to demonstrate that most of the cross-sectional relationship between test scores and suspensions is explained by fixed differences between schools and students. Nonetheless, being suspended predicts lower than *usual* test scores for the affected student, holding fixed her grade and the school she attends. These negative conditional relationships are stronger for higher-level suspensions. Of course, even these conditional relationships may be plagued by selection concerns, which motivate the empirical designs that we focus on in this paper.

4 Natural Experiment: Relaxing School Discipline

To estimate the causal effect of reducing suspension use in New York City schools, we exploit a natural experiment induced by the 2012 discipline reform. We show that average test scores increased for those more affected by the reform, relative to those who were less affected. These benefits were shared across a wide range of students, including those who were unlikely to have been suspended themselves. Our results are robust to alterna-

tive specifications, including different ways of defining treatment, controlling for separate linear trends by treatment group, and balancing treatment groups on demographics.

4.1 Empirical Strategy

The reform we study prohibited schools from suspending students for single instances of disorderly behavior, which are classified as Level 2 infractions, starting with the 2012 school year. At the same time, new language was added to the discipline code to encourage specific alternatives. First, light-touch interventions were encouraged early in the disciplinary process (progressive discipline), with the aim of preventing more serious misbehavior in the future. Second, "restorative interventions" were promoted. These non-punitive interventions focus on improving behavior by building relationships between students, or between students and teachers, often through peer mediation.

How schools responded to the reform

Although schools were encouraged to shift towards more progressive discipline, conversations with school personnel suggest a combination of responses to the reform.[16] This is important to bear in mind when interpreting our results, especially when considering how they might generalize to other settings.

First, schools turned to less disruptive punishments for disorderly behavior, such as removal from a single class. This substitution from missing *all* classes to missing *only one* class prevents one bad student-teacher relationship from disrupting learning across the board. Second, the policy change signaled that the district was serious about reducing its reliance on suspensions, which caused some teachers and administrators to rethink their interactions with students more broadly. For example, they may have become more likely to handle conflict within the classroom rather than escalating it.

Third, schools incorporated more restorative interventions. However, adoption of such interventions was limited at first because they require time and training, which may explain why the effects we see on achievement below are gradual. We do not observe restorative interventions due to a lack of any requirement to report them, but there is evidence linking them to improvements in school climate (Augustine *et al.*, 2018; Adukia *et al.*, 2023). In contrast to our results, Augustine *et al.* (2018) find no corresponding math impact for the elementary school sample where they see a decline in suspension use. However, a potential reason for this is the relatively short follow-up period, which would not have been adequate to detect the effects we see in New York City.[17]

[16] Schools could also have changed their reporting due to the reform. We rule out specific versions of this below, such as systematic upgrades of offenses to Level 3 so that a suspension could be applied.

[17] In their middle school sample, Augustine *et al.* (2018) see a decline in math scores. However, there is no

Identifying variation

We argue that the marked drop in suspension rates induced by the reform is plausibly exogenous because of the sharp timing of the change. As Figure 2 shows, suspensions for Level 2 infractions dropped to zero in 2012. The suspension rate for all infractions fell by 22 percent, although the non-Level-2 suspension rate had been slowly declining for several years prior to the reform.[18] Our conversations with school personnel suggest that the reform was unanticipated, and our analysis shows no evidence of anticipatory effects; in fact, suspensions for Level 2 infractions were at a five-year high in 2011.[19]

Our empirical strategy hinges on the fact that the 2012 policy change affected some students more than others. This variation allows us to compare the changes in test scores between groups of students that were affected to different degrees in a difference-in-differences framework. Specifically, the impact of the policy change depended on the extent to which each student's school used suspensions as a punishment for disorderly behavior. A school did not need to make any changes if it had always relied on other tools such as classroom removals or loss of access to extracurricular activities. But a school that had relied heavily on principal's suspensions was forced to find alternatives.

Definition of treatment groups

Ideally, we would like to know the true extent to which students in each school-grade cell are treated, which would be measured by the policy-induced reduction in their suspension rate. This is not observable, but we can estimate it. Because Level 2 suspension rates fall to zero in 2012, the policy-induced reduction is equal to the suspension rate that would have prevailed absent the reform (s_{jt}^{L2}). We approximate this counterfactual suspension rate by using the average suspension rate during the 2006 and 2007 school years ($\hat{s}_{j}^{L2}$). These early years are then excluded from our estimation sample.

To efficiently capture variation in the impact of the reform, we use these historical suspension rates to define two groups: one with students who are *more affected*, and one with students who are *less affected*. The "High Treatment" group contains students in school-grades with historical suspension rate ($\hat{s}_{j}^{L2}$) greater than the median, who were therefore subjected to a large reduction in suspension use. The "Low Treatment" contains the remaining students, for whom we predict suspensions would have been uncommon even without the reform. Discretization in this manner simplifies both the analysis and communication, and avoids having to assume that the effect of treatment is directly proportional

fall in suspension use in middle school, which suggests that this is not due to reduced suspension use.

[18] We discuss how the reform affected non-Level-2 suspensions in Appendix B.

[19] Some teachers were involved in lobbying efforts that led to the reform, but the proposed changes to the discipline code were not announced until June 2012 – too late to affect suspensions in the 2011 school year.

to $\hat{s}_j^{L2}$. The main reason that we define treatment at the school-grade level is that both principals and teachers play important roles in student discipline.[20]

Our treatment group definitions capture meaningful variation in policy impact. In Panel A of Figure 4, we show the evolution of the Level 2 suspension rate, s_{jt}^{L2}, in each group. The initial gap in suspension rates is around one per 100 students. The gap then narrows initially due to mean reversion, before stabilizing at 0.6 per 100 students. This initial mean reversion motivates our use of 2006 and 2007 to define treatment, while omitting this period from the analysis of outcomes. In the 2012 school year, the Level 2 suspension rate falls sharply to zero, implying a policy impact that is nearly three times larger for the High than the Low Treatment group.

The impact of the policy was not necessarily limited to suspensions for Level 2 infractions. Panel B of Figure 4 shows that total suspensions dropped sharply in the High Treatment group between 2011 and 2012, separate from the downward secular trend throughout the sample period.[21] The magnitude of this relative fall is 2.5 suspensions per 100 students, which is more than can be explained by the elimination of Level 2 suspensions. Figure 2 shows that there was a fall in Level 3 suspensions at the time of the reform. This is not surprising: Responses by educators could have affected suspensions for Level 3-5 infractions as well, as we discuss in Appendix B. Indeed, our conversations with stakeholders suggested that the reform helped trigger a broader rethink of disciplinary approaches.

Our results are qualitatively robust to specifying treatment in many alternative ways, as we explore in Appendix E. First, in Appendix E.2, we use $\hat{s}_j^{L2}$ as a continuous measure of treatment. Second, in Appendix E.3, we define our treatment groups using the full preperiod rather than only 2006 and 2007. Third, Appendix E.4 generalizes to a larger number of discrete treatment groups, with larger impacts for those who are more affected. Finally, Appendix E.5 shows results when treatment is defined using total suspension rates.

Because students are not randomly assigned to schools, there are some differences between the two treatment groups, but the differences are smaller than one might have expected from the difference in suspension rates. Table 2 presents baseline characteristics by group. Students in the High Treatment group have lower test scores than those in the Low Treatment group by about 0.04 standard deviations, and higher suspension rates for all types of infractions. The remaining differences are quite small: High Treatment school-grades are larger, have slightly higher shares of white and slightly lower shares of black and Hispanic students, and have relatively more 8th than 6th graders. Although balance

[20] Teachers have discretion over when to escalate misbehavior, but final decisions are made by principals. Additionally, teachers affect student behavior through everyday interactions in class.

[21] Relative to the Level 2 infractions in Panel A, mean reversion of total suspension rates in Panel B is more gradual. However, it still appears to be complete by the time the reform is implemented.

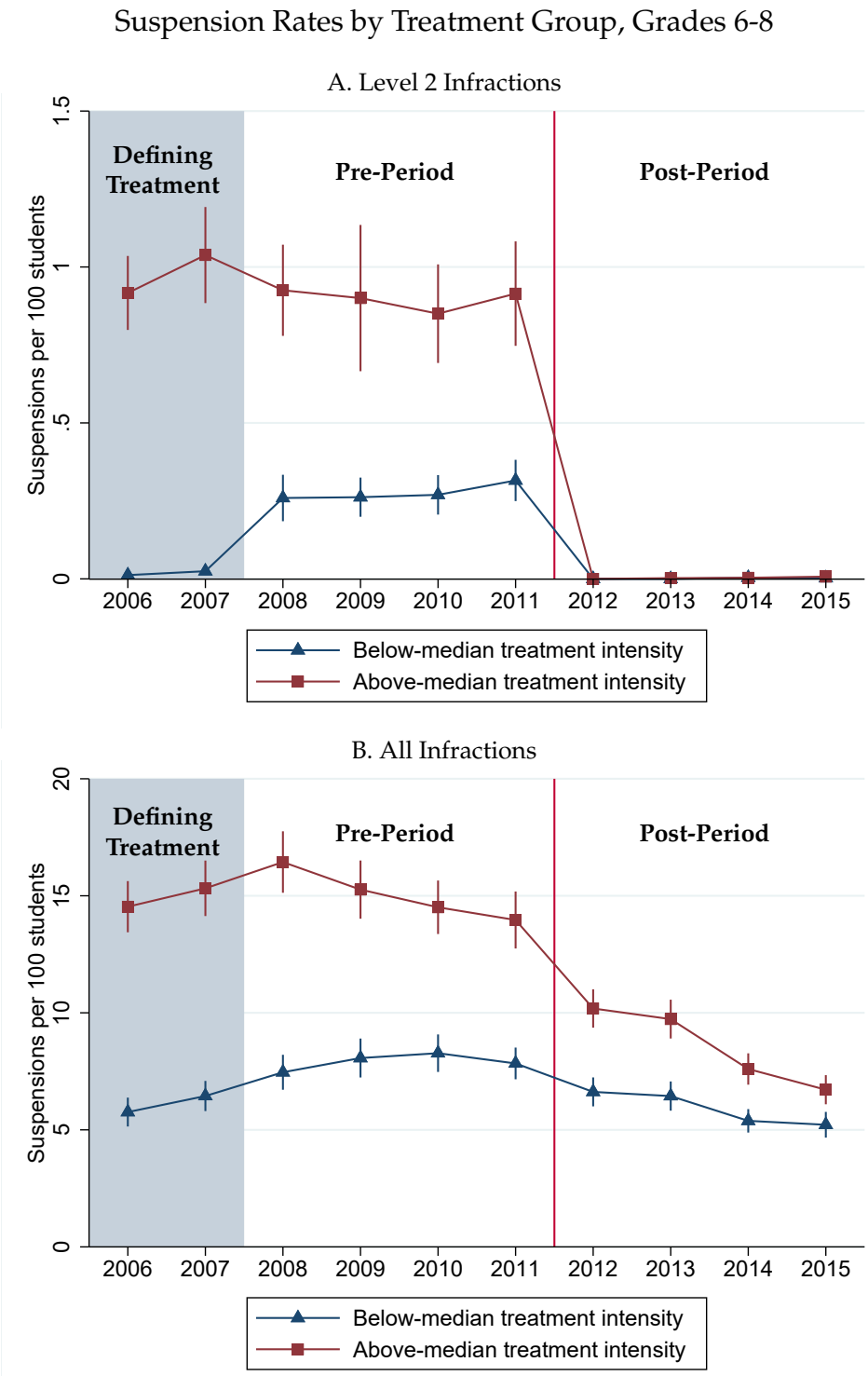

Figure 4. This figure shows trends in suspension rates for our below-median treatment intensity group (Low Treatment) and our above-median treatment intensity group (High Treatment). Panel A plots suspension rates for Level 2 infractions only; Panel B plots suspension rates for all infractions. The shaded years are used to assign treatment, and are excluded from our estimates of treatment effects. The vertical red line shows the timing of the reform. Figure I5 provides a regression-adjusted version of Panel A. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education.

Table 2. Pre-Period Treatment Group Characteristics, Grades 6-8

Variable	Low Treatment	High Treatment	Difference
Test Scores (Math)	0.021	-0.020	(0.261)
Test Scores (Reading)	0.019	-0.018	(0.253)
% Black	0.281	0.265	(0.446)
% Hispanic	0.423	0.389	(0.092)
% White	0.125	0.171	(0.006)
% Male	0.506	0.513	(0.011)
% Free Lunch	0.672	0.669	(0.875)
Susp. / 100 (Level 2)	0.277	0.898	(0.000)
Susp. / 100 (Level 3)	1.842	5.094	(0.000)
Susp. / 100 (Level 4)	4.827	7.996	(0.000)
Susp. / 100 (Level 5)	0.965	1.075	(0.115)
% Grade 6	0.382	0.265	(0.001)
% Grade 7	0.316	0.362	(0.223)
% Grade 8	0.302	0.373	(0.064)
Expenditure / Student	17,141	17,340	(0.322)
Teacher Salary	70,608	71,896	(0.000)
Avg. School-Grade Size	257.0	317.8	(0.000)
Students per Teacher	14.6	14.4	(0.121)
Students per Counselor	293.5	287.2	(0.543)

Table notes. This table compares the High and Low Treatment groups on a number of pre-reform characteristics, averaged over 2008-2011. The third column contains p-values from pairwise comparisons between groups with standard errors clustered at the school-grade level. Test scores are standardized within the sample in subject-grade-year cells. Suspension rates are expressed as the number of suspensions per 100 students. Data are from the New York City Department of Education, and include students in grades 6-8.

is not required for identification, we show in Appendix E.1 that reweighting to balance on these covariates has very little effect on our results.

Estimating treatment effects

Using the variation provided by the 2012 reform and two-way fixed effects, we estimate the effect of reducing schools' reliance on suspensions in a difference-in-differences framework.[22] Our analysis relies on the strong parallel trends assumption: i.e., The average change in test scores in school-grades if they had faced a given level of exposure to the

[22]There is a rapidly evolving literature on problems that can arise with two-way fixed effects regressions in some settings (see e.g., de Chaisemartin and D'Haultfœuille (2022)). We avoid problems with "negative weights" because our treatment is not staggered. However, some weighting issues also arise with time-varying controls. For this reason, we replicate our analysis without any time-varying controls. Figure I17 shows the results, which are nearly identical to our main figures.

reform equals the average change for school-grades that were in fact exposed to that degree.[23] Importantly, there is no need to assume that changes in suspension rates were randomly assigned. This would be unreasonable, since pre-period suspension rates could be associated with characteristics of schools, teachers, students, and administrators.

For each outcome y_{ijt}, we estimate Equation 1 using data from 2008 to 2015.

$$y_{ijt} = \underbrace{\alpha_j + \gamma_t}_{\textbf{Fixed effects}} + \sum_{k \neq 2011} \rho_k \Big[\underbrace{\mathbb{1}(t = k)}_{\textbf{Time}} \times \underbrace{\mathbb{1}(\hat{s}_j^{L2} \geq \text{Median}(\hat{s}_j^{L2}))}_{\textbf{High Treatment}} \Big] + \underbrace{\beta X_{ijt}}_{\textbf{Controls}} + \varepsilon_{ijt} \quad (1)$$

In this equation, α_j and γ_t are school-grade and year fixed effects, and X_{ijt} includes race, gender, English Language Learner status, and free or reduced price lunch status. Our primary outcomes of interest are annual student exam scores in math and reading.

The coefficients, ρ_k, measure the difference in the gap between the High Treatment and Low Treatment groups in each year, relative to the gap in 2011. Prior to the reform, we would hope to (and do) see parallel trends for the two groups, as reflected by $\rho_k \approx 0$ in all pre-period years. Then, if relaxing the discipline policy had a positive causal effect on average achievement, ρ_k would rise from 2012 onward as the test scores of the High Treatment and Low Treatment groups converge toward each other. Conversely, a negative causal effect would lead to diverging test scores and falling values of ρ_k over time.

The most important threat to our identification is that schools implemented other contemporaneous policy changes – or were hit by contemporaneous shocks – that differentially affected schools with higher versus lower suspension rates. We discuss briefly below why it is unlikely that any such violations of our parallel trends assumption are driving our results. A more detailed discussion is available in Appendix F.

4.2 Impact on Average Achievement

Our results suggest that eliminating suspensions for Level 2 infractions was, on average, beneficial for student achievement in New York City schools. In fact, the reform entirely closed the gap in math test scores between the High and Low Treatment groups.

Figure 5 shows how math scores evolve in each treatment group. The top panel shows the average levels of standardized scores in the two groups. Because test scores are standardized within our sample of New York City public school students, changes in the two groups are relative to each other.[24] The bottom panel shows estimated treatment effects

[23] As Callaway *et al.* (2021) discuss, this is stronger than the standard parallel trends assumption, which is that test scores in the two groups would have moved together absent the reform. However, the assumptions are equivalent if treatment effects are not heterogeneous. We note that we find very little evidence of heterogeneity in treatment effects on observables.

[24] In Appendix C, we re-normalize scores to all of New York State. Figure C1 shows that the relative gains

(ρ_k) from Equation 1, with 95 percent confidence intervals. Prior to the 2012 reform, average test scores are roughly 0.04 standard deviations higher in the Low Treatment group than in the High Treatment group. The average scores of the two groups evolve in parallel, and ρ_k is not statistically different from zero in any pre-treatment year.

Following the 2012 reform, there was a substantial relative improvement in High Treatment school-grades. Gains accumulated gradually, reaching a maximum of 0.05 standard deviations in 2014. At this point, the original gap in test scores between the High and Low Treatment groups had been entirely closed. The gradual improvements we see are consistent with our results in Section 5, which suggest an important role for school culture. It may take time for such cultural change to occur, and then to translate into test score improvements (Anfara Jr. *et al.*, 2013).

Our estimated treatment effects are large, especially given the simplicity of the reform and its low financial cost. The gains are equivalent to improving teacher value-added by one third of a standard deviation (Chetty *et al.*, 2014). They are also about one quarter of the benefit of attending a smaller class in grades K-3 under Tennessee's Project STAR (Krueger, 1999), but Krueger estimated that implementing these class size reductions more widely would have cost $2,151 per student for each of the four years of the intervention.

Table 3 shows that our results are robust to several generalizations of our approach. Column 1 shows our baseline yearly estimates from Equation 1. In Column 2, we remove student demographic controls. In Column 3, we add these back along with controls for baseline math and reading test scores from grade 3 as a noisy measure of student ability.[25] In Column 4, we allow the two treatment groups to follow distinct linear trends. In Column 5, we add controls for suspensions for higher-level infractions that were not directly targeted by the reform.

Figure 6 shows that reading scores follow a similar pattern, but with smaller treatment effects. Columns 6-10 of Table 3 show that the statistical significance levels of the reading estimates are thus more fragile. Controlling for separate linear trends for each treatment group further reduces the size of the reading estimates: in that specification, we cannot conclude that the effects are different from those for math, but neither can we rule out the possibility that the reform had no impact on reading scores. This stems from a gradual increase in the reading scores of the High Treatment group relative to the Low Treatment group prior to the reform, although no pre-period coefficient is statistically significant.

are largely driven by an improvement in the High Treatment group relative to students outside of New York City, with some evidence of a small improvement for the Low Treatment Group. In addition, data on student behavior suggest that the change was in High Treatment schools (see Section 5).

[25] We use grade 3 scores to avoid controls that are endogenous to the reform. Note that only 6th graders in 2015 have grade 3 scores from 2012 or later. Our results are robust to the exclusion of these students.

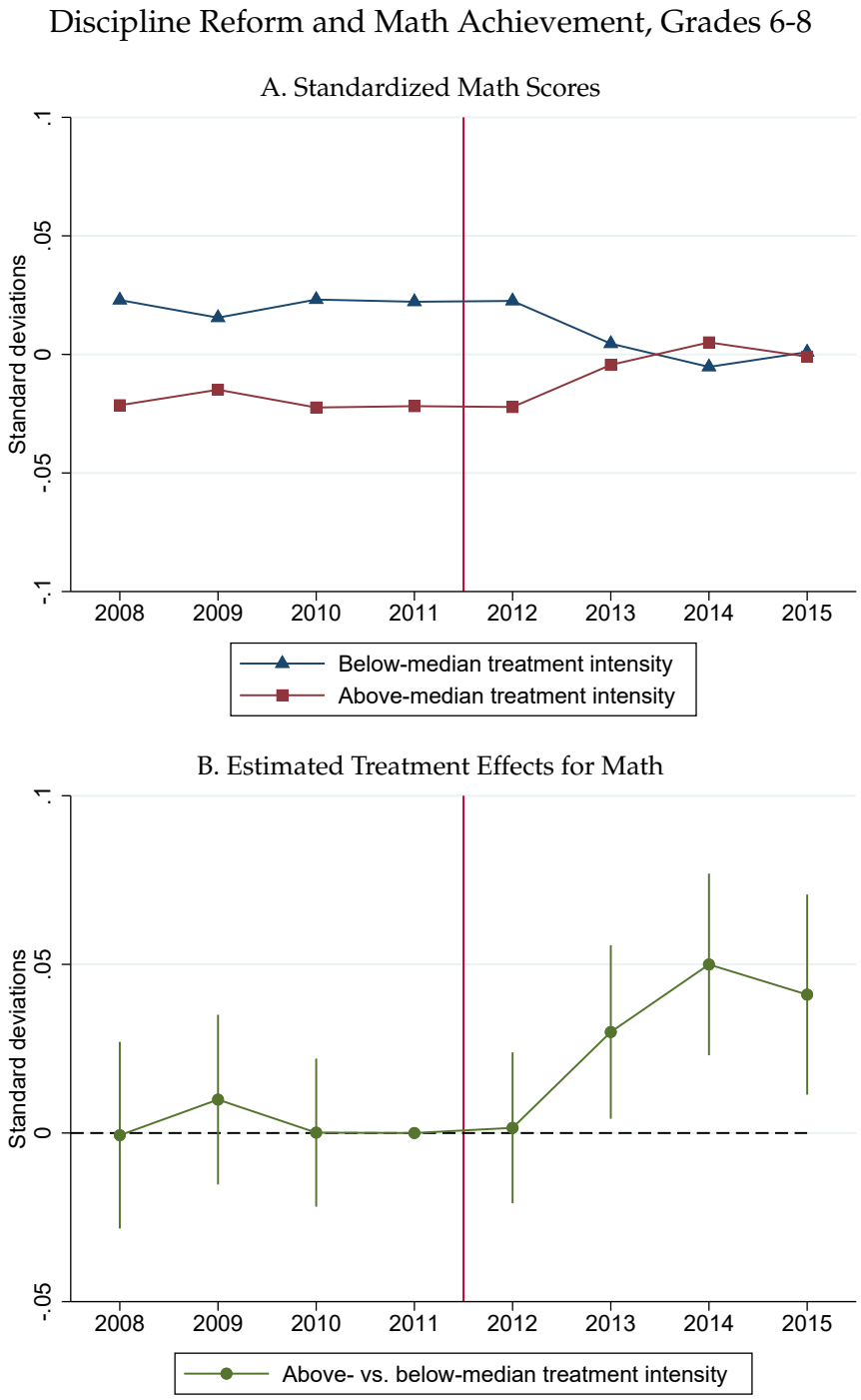

Figure 5. This figure shows the effects of the 2012 reform on math achievement. Panel A plots average standardized math scores in each treatment group over time. Panel B plots the estimated treatment effects, ρ_k, from Equation 1. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

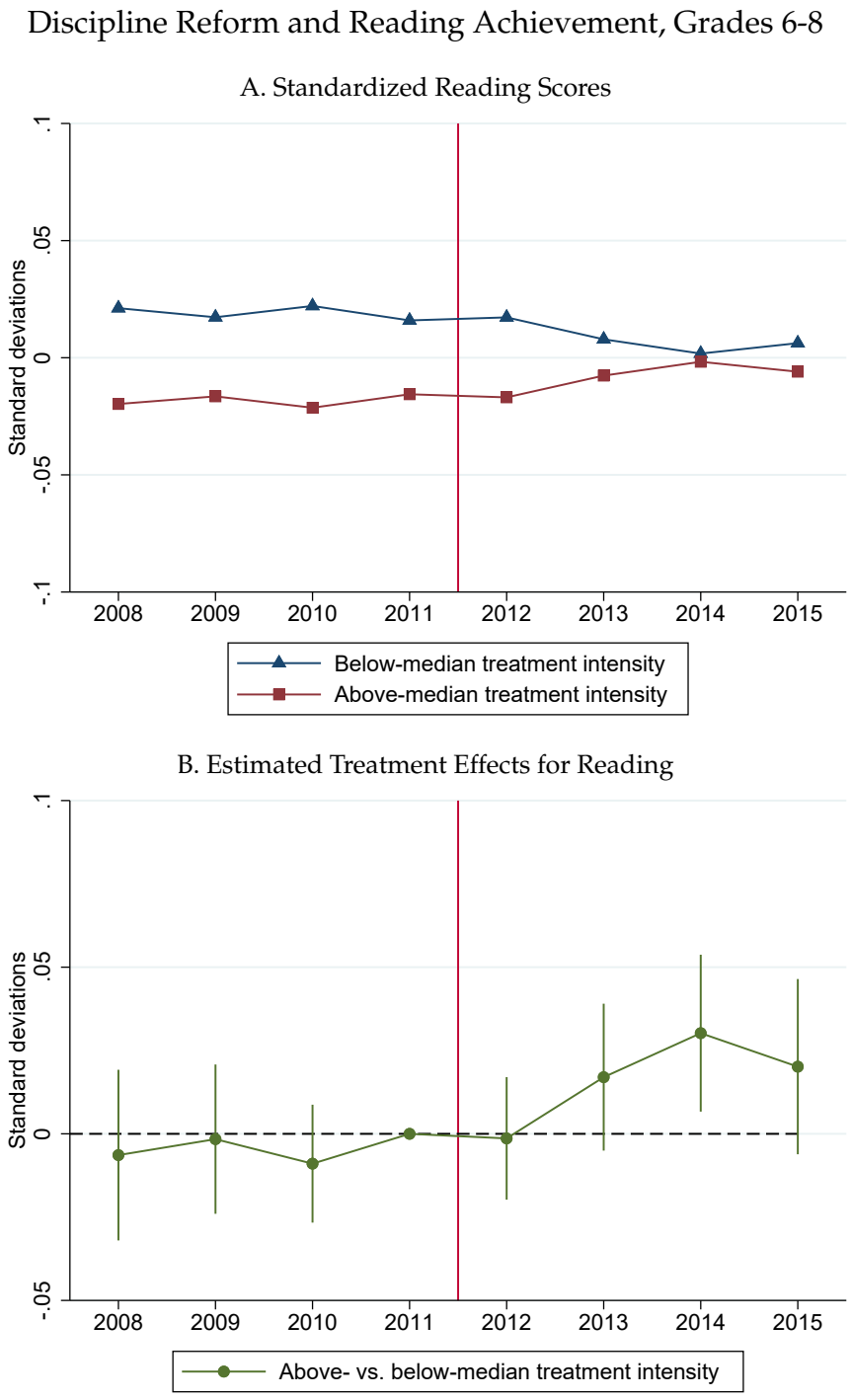

Figure 6. This figure shows the effects of the 2012 reform on reading achievement. Panel A plots average standardized reading scores in each treatment group over time. Panel B plots the estimated treatment effects, ρ_k, from Equation 1. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Table 3. Estimated Treatment Effects on Standardized Test Scores, Grades 6-8

	(1) Math	(2) Math	(3) Math	(4) Math	(5) Math	(6) Reading	(7) Reading	(8) Reading	(9) Reading	(10) Reading
ρ^{2012}	0.00153	0.00390	0.00461	0.00110	-0.00228	-0.00138	0.00178	-0.00183	-0.0000922	-0.00438
	(0.0114)	(0.0119)	(0.0112)	(0.0130)	(0.0113)	(0.00937)	(0.0101)	(0.00889)	(0.0115)	(0.00929)
ρ^{2013}	0.0299	0.0337	0.0362	0.0303	0.0260	0.0170	0.0210	0.0196	0.0171	0.0138
	(0.0131)	(0.0136)	(0.0129)	(0.0175)	(0.0130)	(0.0112)	(0.0120)	(0.00991)	(0.0162)	(0.0112)
ρ^{2014}	0.0500	0.0553	0.0555	0.0511	0.0435	0.0302	0.0338	0.0346	0.0291	0.0250
	(0.0137)	(0.0145)	(0.0132)	(0.0210)	(0.0136)	(0.0120)	(0.0131)	(0.0106)	(0.0198)	(0.0119)
ρ^{2015}	0.0410	0.0449	0.0446	0.0429	0.0331	0.0201	0.0243	0.0220	0.0179	0.0140
	(0.0151)	(0.0160)	(0.0139)	(0.0250)	(0.0150)	(0.0134)	(0.0148)	(0.0116)	(0.0245)	(0.0134)
Stud. Dem.	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Prior Test Scores	No	No	Yes	No	No	No	No	Yes	No	No
Time Trends	No	No	No	Yes	No	No	No	No	Yes	No
Other Susp.	No	No	No	No	Yes	No	No	No	No	Yes
Observations	1350606	1350606	1063802	1350606	1350606	1325715	1325715	1067321	1325715	1325715
Clusters	1057	1057	1055	1057	1057	1057	1057	1055	1057	1057

Table notes. Columns (1) and (6) of this table show estimated treatment effects, ρ_k, from Equation 1. The dependent variable is a student's standardized test score in math or reading. Controls include demographics and fixed effects for year and school-grade. Columns (2) and (7) remove student demographic controls, leaving only the two-way fixed effects. Columns (3) and (8) add controls for student test scores in grade 3. Columns (4) and (9) include time trends by treatment group. Columns (5) and (10) include controls for non-Level-2 suspensions. Standard errors in all specifications are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

The pattern of test score gains we observe in math and reading cannot easily be explained by other contemporaneous policy changes. In Section 6, we rule out policies that provide additional staffing or funding to high-suspension schools. In Appendix F, we then discuss why the timing of the test score effects we see in Figures 5 and 6 cannot be explained by the revision of state assessments to align with the Common Core in 2012, the election of Mayor Bill de Blasio, or his appointment of Chancellor Carmen Fariña. In our conversations with school administrators and staff, we have not become aware of any other policy change that could drive our effects. Furthermore, since the High Treatment group has slightly fewer minority students than the Low Treatment group (see Table 2), policies aimed at closing the achievement gap would bias *against* our treatment effects.

Absolute Effects Of The Reform

Our results show that test scores improved for the High Treatment group relative to the Low Treatment group. However, the Low Treatment group is also somewhat affected by the reform. One way to think about this is to infer the impact on the Low Treatment group by making the strong assumption that effects are linear in the reduction in suspensions. Linearity implies a test score impact of 0.025 standard deviations in 2014 for the Low Treatment Group. In turn, this implies that the total impact on the High Treatment group (compared to no reform) is 0.075 standard deviations, and that the average impact across all of New York City is around 0.05 standard deviations. As it turns out, this is roughly equal to the impact on the High Treatment group relative to the Low Treatment group.

We provide an alternative way of thinking about the overall impact in Appendix C, where we re-normalize test scores to be relative to all students in the state, rather than only the city. Students outside of the city were not subject to this reform, so the analysis provides a suggestive indication of the absolute impact. The results there also suggest that the High Treatment group saw a large improvement relative to other students in the state, with some indication of a small positive effect on the Low Treatment group.

4.3 Treatment Effect Heterogeneity

We next show that the benefits of the 2012 discipline reform were broad-based. Students benefited even if they would have been unlikely to be suspended themselves.

We start by estimating an individual's likelihood of receiving any suspension (their "suspension risk"). To do so, we estimate the relationship between individual characteristics and suspension outcomes in 2006-2007. Then we use those estimated parameters to generate predictions for 2008-2015.[26]

[26] Appendix Figure I7 shows results from the same exercise with predictions for Level 2 suspensions.

Table 4. Pre-Period Characteristics by Suspension Risk Quartile, Grades 6-8

	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Math Scores	0.814	0.093	-0.147	-0.568
Reading Scores	0.782	0.162	-0.096	-0.517
% Black	0.021	0.219	0.412	0.621
% Hispanic	0.189	0.546	0.451	0.351
% White	0.261	0.189	0.129	0.021
% Male	0.274	0.233	0.551	0.938
% Free Lunch	0.456	0.662	0.734	0.834
Susp. / 100 (All)	2.739	7.676	14.751	25.844
Susp. / 100 (Level 2)	0.131	0.379	0.787	1.324
Total Students	430,994	431,138	431,621	431,618

Table notes. This table shows characteristics by suspension risk quartile, averaged over 2008-2011. Quartile 1 contains students with the lowest suspension risk, Quartile 4 the highest. Test scores are standardized within subject-grade-year cells. Suspension rates are expressed as the number of suspensions per 100 students. Data are from the New York City Department of Education, and include students from grades 6 through 8.

We model the probability that student i received at least one suspension (i.e., $S_{ijt} > 0$) in year t using a logit specification.[27]

$$\Pr(S_{ijt} > 0 \mid X_{ijt}) = \frac{1}{1 + e^{-(\alpha + \beta X_{ijt} + \epsilon_{ijt})}} \tag{2}$$

The matrix X_{ijt} includes: (1) demographics (race, gender, English Language Learner, free lunch status); and (2) grade 3 test scores for math and reading as proxies for ability.

Table 4 shows summary statistics by suspension risk quartile prior to the reform. Students in the lowest risk quartile (Quartile 1) score 1.4 standard deviations higher on their math exams than students in the highest quartile (Quartile 4), and 1.3 standard deviations higher in reading. Similarly, 21 percent of students in Quartile 1 are black or Hispanic, compared to 97 percent in Quartile 4. Actual suspension rates are 10 times higher in Quartile 4 than in Quartile 1, both for all suspensions and Level 2 suspensions.

Estimating treatment effects by quartile of suspension risk

Next, we estimate Equation 1 by quartile of suspension risk. Figure 7 plots the treatment effect for each quartile in 2014 (ρ_{2014}) with 95 percent confidence intervals. These coefficients measure how the test score gap between High and Low Treatment groups changed between 2011 and 2014. For comparison, our estimates for the full sample – 0.05σ for math and 0.03σ for reading – are shown as red dashed lines. The green line plots the number of

[27] We chose a binary specification because most students are suspended either one or zero times. The results are extremely similar if we use a Poisson count specification.

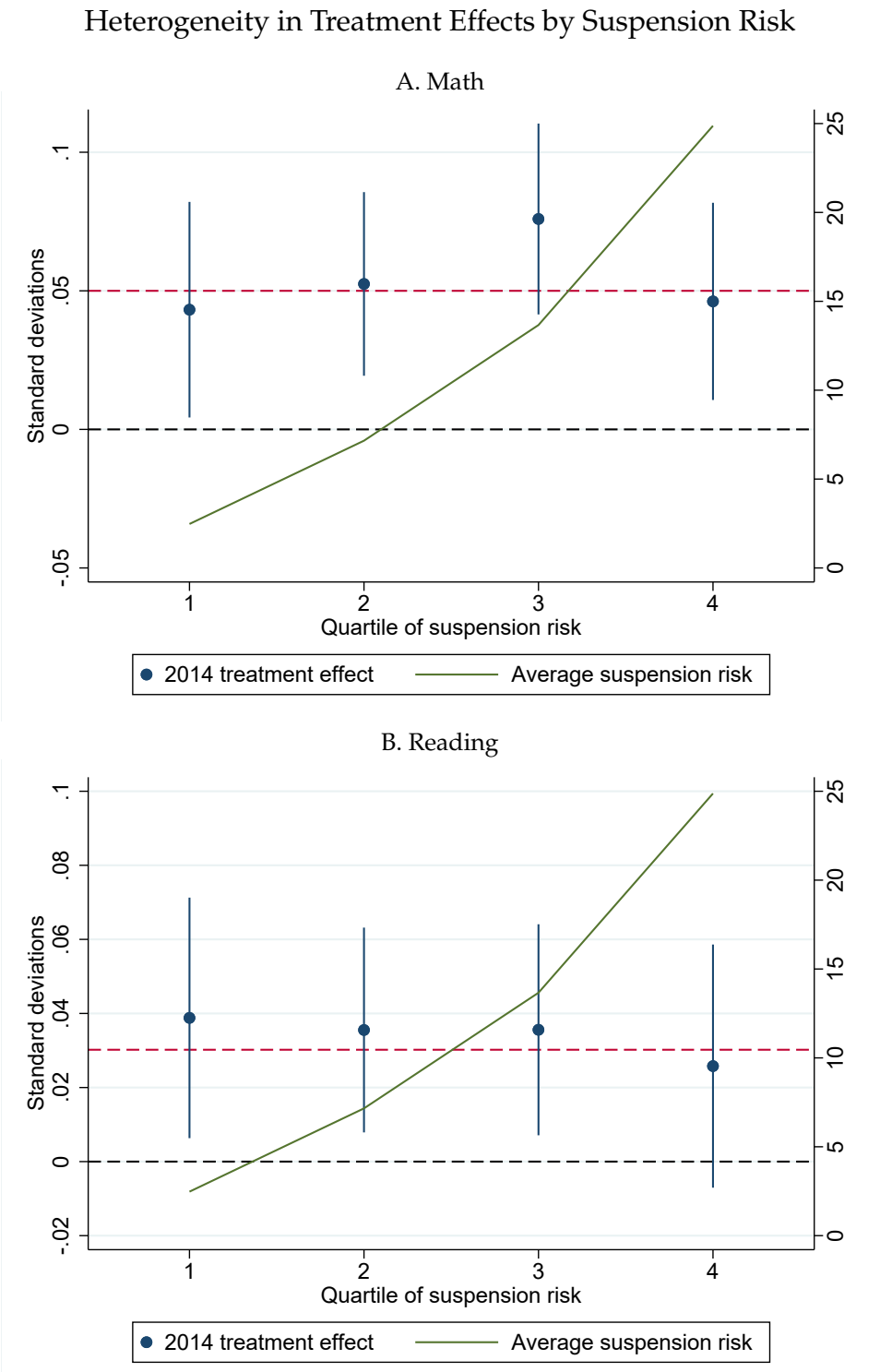

Figure 7. This figure shows the effects of the 2012 discipline reform on test scores for students in different quartiles of predicted suspension risk. Panel A plots the estimated treatment effects on math scores in each risk quartile in 2014 from Equation 1. Panel B plots estimated treatment effects on reading scores in each risk quartile in 2014. The vertical blue lines show 95 percent confidence intervals. The red dashed line shows the estimated treatment effect in 2014 for the full sample. The green line shows actual suspension rates for all infractions within each quartile over the 2008-2011 period. Data are from the New York City Department of Education, and include students from grades 6 through 8.

suspensions of any type per 100 students in each quartile prior to the reform. Event study graphs for each quartile are available as Appendix Figures I8 and I9.

Our results suggest broad-based gains: The gradient of treatment effects by suspension likelihood is relatively flat for both math and reading, although we lack the power to rule out small differences across quartiles. Some of the smallest point estimates for both subjects are in the highest risk quartile, despite suspension rates increasing sharply across quartiles. Statistical imprecision makes us reluctant to draw strong conclusions, but it is plausible that the disadvantaged and underperforming students who are most at risk of suspension are not well-positioned to benefit from the changes brought by the reform.

In Appendix D, we also estimate Equation 1 for subgroups who are more or less at risk of suspension on average. Notably, we find nearly identical treatment effects for boys and girls despite boys being suspended twice as often as girls prior to the reform.

5 Impact of the 2012 Reform on School Culture

The results so far have established that the 2012 relaxation of school discipline led to achievement gains that benefited a broad range of students, including those who would not likely have been suspended themselves. We next present evidence that improvements in school culture are responsible for these gains. This helps explain why the benefits of the reform too time to be realized. The pattern of results is also consistent with public arguments in favor of relaxing school discipline (Mukherjee, 2007; Miller *et al.*, 2011), our conversations with teachers and reform advocates, and academic literature which links school climate to academic achievement (Dulay and Karadağ, 2017; Pas *et al.*, 2019).

To measure school culture, we use data from student and teacher surveys. We then estimate treatment effects on these measures in the same way that we did for achievement. Appendix Table I3 shows the precise questions we analyze. They relate to student-teacher relationships, behavior, or feelings of safety at school.

Student-teacher relationships

As a proxy for the quality of student-teacher relationships, we look at student and teacher responses to the question of whether "most students at my school treat adults with respect." Responses are on a four-point scale from "strongly agree" to "strongly disagree."[28] We code these responses so that higher is better. We then standardize the numeric responses within each year, and report results in standard deviation units.

[28]From 2006 to 2011, the student (but not the teacher) question contained a "don't know" option. We exclude these responses. The share of students selecting the "don't know" option is roughly 7 percent in both the High and Low Treatment groups. Other culture measures have consistent options over time.

Perceptions of safety and behavior

We construct two indices to measure how safe students and teachers feel: one directly measures safety, and the other measures student behavior. The behavior questions ask about the frequency of incidents such as bullying and fighting. Answers are on a four-point scale from "none of the time" to "all of the time." The safety questions ask students whether they feel safe in areas such as classrooms, bathrooms, and hallways. Teachers are asked about order and discipline at school, and whether crime and violence are a problem. Answers range from "strongly agree" to "strongly disagree." Responses are coded so that higher numbers are better, and standardized by year. To construct each index, we average the standardized responses across questions, and re-standardize the combined index.

Perceptions of bias

We cannot directly measure perceptions of bias, but we find a pattern consistent with reduced bias playing a role. Specifically, Appendix Figure I10 shows that larger reductions in racial disparities in suspension rates between minority and white students are associated with larger average achievement gains across all students.

This is consistent with qualitative evidence elsewhere, which suggests that students perceive disciplinary environments to be unfair if rules are applied inconsistently, or if there are severe penalties for trivial infractions (Morrison, 2018). Indeed, suspension for the types of low level infractions that were the focus of this reform were also likely to be relatively discretionary, which makes it more likely that they were perceived to be applied unequally. Such perceptions of unfairness have been linked to less positive student-teacher relationships and worse behavior (Way, 2011), which may be part of the reason why their abolition leads to improvements on these dimensions in New York City.

Estimating treatment effects

For each culture measure, we estimate Equation 3. This is analogous to Equation 1. However, we cluster standard errors by school because outcomes are measured at the that level. We do not include individual individual-level controls.[29]

$$y_{ijt} = \underbrace{\alpha_j + \gamma_t}_{\textbf{Fixed effects}} + \sum_{k \neq 2011} \rho_k \Big[\underbrace{\mathbb{1}(t = k)}_{\textbf{Time}} \times \underbrace{\mathbb{1}(\hat{s}_j^{L2} \geq \text{Median}(\hat{s}_j^{L2}))}_{\textbf{High treatment}} \Big] + \varepsilon_{ijt} \tag{3}$$

Just as with test scores, treatment effects are given by the yearly ρ_k coefficients. These can be interpreted as the difference on a given culture metric between the High and Low Treatment groups relative to the difference in 2011, just prior to the reform.

[29] We retain fixed effects at the school-grade level, which is possible because grades within schools shift in relative size over time. Little changes with school instead of school-grade fixed effects.

5.1 Treatment Effects on Culture

Figure 8 (Panels A-C) and Table 5 display our estimates of ρ_k in Equation 3 for each measure of school culture.[30] We find positive effects on both student and teacher assessments of student-teacher relationships, feelings of personal safety, and perceptions of student behavior. Prior to the reform, the average levels of all the culture measures evolve largely in parallel for the High and Low Treatment groups, and ρ_k is never statistically different from zero for any measure. All measures then show a relative improvement in the High Treatment group after the reform was implemented in 2012. The gains reported by teachers are generally larger, and eventuate more quickly, than those for students.[31]

Table 5. Estimated Treatment Effects on School Culture Outcomes, Grades 6-8

	(1) Respect (Students)	(2) Respect (Teachers)	(3) Behavior (Students)	(4) Behavior (Teachers)	(5) Safety (Students)	(6) Safety (Teachers)
ρ^{2012}	0.0613	0.154	0.0169	0.133	0.0767	0.119
	(0.0581)	(0.0797)	(0.0490)	(0.0681)	(0.0534)	(0.0801)
ρ^{2013}	0.0640	0.144	0.0366	0.161	0.0743	0.156
	(0.0678)	(0.0845)	(0.0623)	(0.0720)	(0.0661)	(0.0774)
ρ^{2014}			0.0303		0.132	
			(0.0720)		(0.0669)	
ρ^{2015}			0.150		0.186	
			(0.0746)		(0.0824)	
Observations	1023406	1024833	1321664	1024833	1321664	1024833
Clusters	392	392	392	392	392	392

Table notes. This table shows show estimated treatment effects ρ_k from Equation 3 on our culture outcomes. Controls include year and school fixed effects. Standard errors are clustered at the school level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Linking culture and test score gains

There are two senses in which these improvements in culture are consistent with the gains in achievement. First, school-grades with larger improvements in culture also saw greater test score gains. Second, the relative improvements in culture and test scores are remarkably consistent with the cross-sectional relationships between these measures.

Panels D to F of Figure 8 compare the change in each culture measure to the math score gain in each school-grade. For all three measures, there is a positive correlation between culture improvements and math score improvements. The same is true for reading (see

[30] Appendix Figure I12 shows the levels of our standardized culture measures in each group over time.

[31] The effects on teacher perceptions precede the test score impacts. This should be expected, since our culture measures reflect perceptions in the spring of each school year, but test scores measure learning over the whole year. Additionally, it may take time for culture improvements to affect achievement.

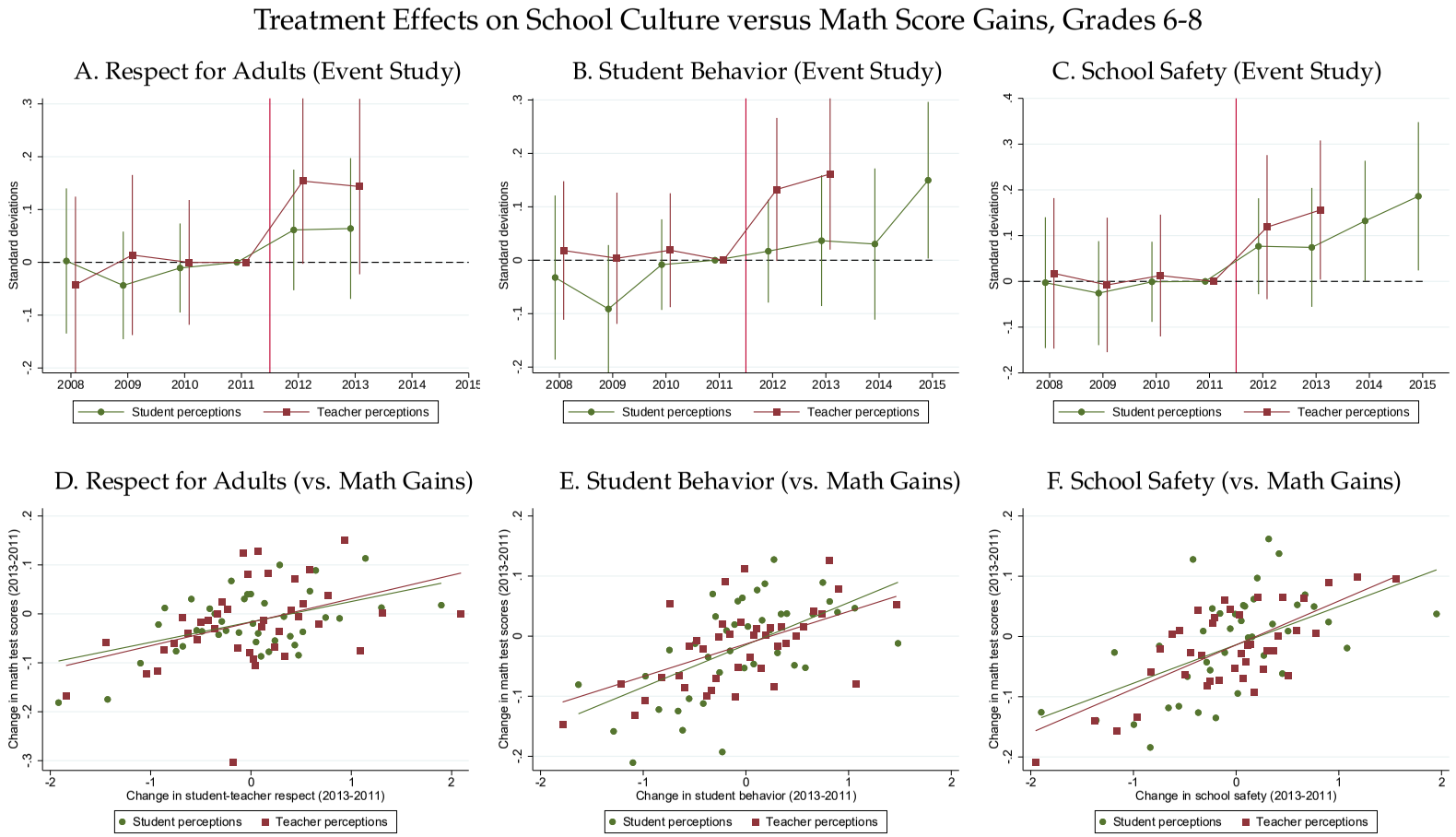

Figure 8. This figure shows the effects of the 2012 discipline reform on our three measures of school culture and the relationship between improvements in culture and test score gains at the school-grade level. Panel A plots the estimated treatment effects ρ_k from Equation 3 on respect, with student responses indicated by green circles and teacher responses by dark red squares. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. In Panel D, we calculate the changes in student and teacher respect between 2011 and 2013 in each school-grade and create binned scatterplots against the corresponding changes in math test scores. Bins of student responses are indicated by green circles and bins of teacher responses by dark red squares. Panels B and E repeat the same analysis for perceptions of student behavior, and Panels C and F do the same for safety. Culture metrics are standardized within each year. Test scores are standardized within the sample in grade-year cells. Standard errors in all regressions are clustered at the school level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Appendix Figure I13). Although our identification strategy does not tell us the nature of the relationship between culture and achievement, these correlations are consistent with school culture being an important driver of the test score gains from the reform. We also note that other studies have suggested that improvements in school climate can raise student achievement (Kraft *et al.*, 2016; Dulay and Karadağ, 2017; Pas *et al.*, 2019).

Comparing the sizes of gains in culture and achievement, the respect measure reported by teachers improved by 0.15 standard deviations in the High Treatment group relative to the Low Treatment group, while the gain in test scores was 0.05 standard deviations. The cross-sectional correlation between our teacher respect measure and math scores is 0.32. Thus, if we were to draw two random students and observed a difference of 0.15 standard deviations in teachers' assessments of respect at their schools, we would indeed predict the test score of the student at the higher-respect school to be 0.05 standard deviations higher. The 0.15 standard deviation improvement in teachers' perceptions of student behavior predicts a similar 0.04 standard deviation math gain.[32]

5.2 Improvements in Measures of Student Behavior

Consistent with teacher and student perceptions, we also find evidence of improvements in actual behavior. Specifically, we see a decline in the number of non-violent, disruptive incidents per student reported via New York State's VADIR system. Figure 9 shows the estimated treatment effects from Equation 3. Panel A shows that disruptive incidents in High Treatment school-grades drop relative to Low Treatment school-grades after the reform, while Panel B quantifies the treatment effects. These results suggest that student behavior improved despite reduced deterrence from the threat of suspension.

We caution that schools may have an incentive to under-report the number of VADIR incidents. This is most likely to be an issue for violent rather than disruptive behavior: Violent incidents in this system are used to calculate a School Violence Index (SVI), and schools may be designated as "persistently dangerous" if they have consistently high SVIs.[33] This can affect funding and enrollment. Although the disruptive incidents that we examine are not included in a school's SVI, schools could conceivably misreport these incidents as well. Of course, any incentive to misreport would have to vary by treatment group and change sharply in 2012 to explain the treatment effects in Figure 9.

[32]Teacher perceptions of behavior improve by 0.15σ. Given a cross-sectional correlation between the behavior index for teachers and math scores of 0.27, this also predicts a 0.04σ higher math score.

[33]We do not find effects on the SVI, which is an average of incidents from homicide and sexual assault to criminal mischief and larceny (see Appendix Figure I16). This may be because violent incidents are rare.

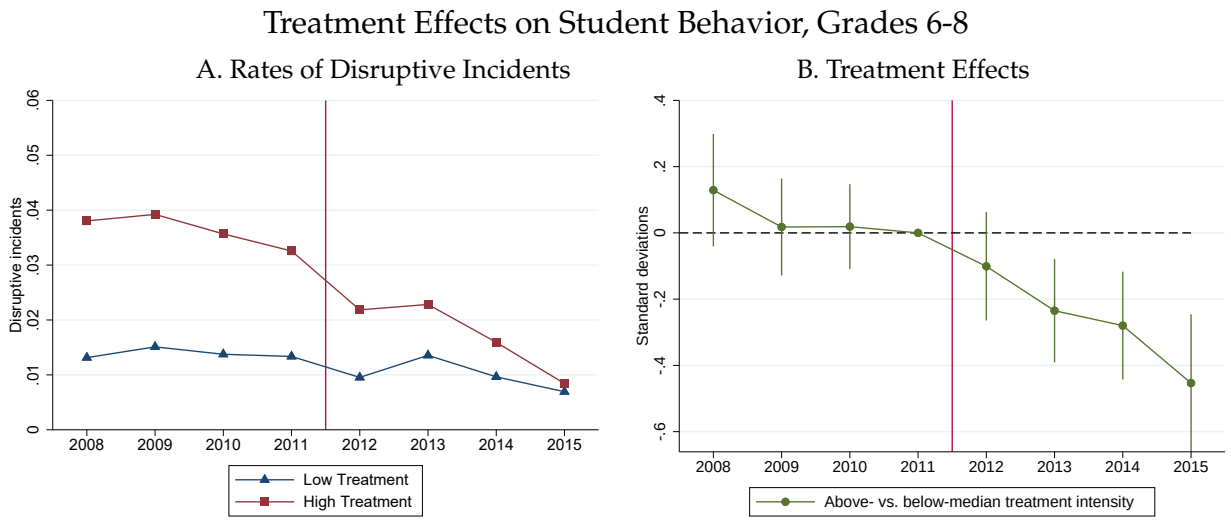

Figure 9. This figure shows the effects of the 2012 reform on disruptive incidents recorded in VADIR. Panel A shows average rates of disruptive incidents per student. Panel B plots estimated treatment effects from Equation 3. Each point measures the gap in the rate of disruptive incidents, in standard deviation units, between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6-8.

6 Ruling out Other Potential Mechanisms

Our final step is to turn our attention to other potential mechanisms. We begin by arguing that at most a very small part of the achievement gains we see can be driven by the direct effects of changing the punishment of students whose behavior would previously have earned a suspension. We then provide suggestive evidence that the benefits of the reform were not driven by staffing or funding changes, or by student sorting. In Appendix G we discuss why changes in teacher composition cannot explain our results either.

6.1 Direct Effects on Suspended Students

A simple calculation reveals that the direct impact of each suspension would have to be extremely large to explain the average gains from this reform. The treatment effect on math scores is 0.05 standard deviations in 2014, while the difference in the decline in Level 2 suspension rates between our treatment groups is 0.6 per 100 students. Thus, if the gains came solely from the elimination of direct effects, the implied impact per suspension is over $0.05/0.006 = 8$ standard deviations. Even if the reform drove the entire reduction in suspensions for *all* infractions between 2011 and 2014, the implied effect per suspension is 1.3 standard deviations. Such gains are implausibly large, given that suspended students perform only 0.6 standard deviations worse *unconditionally* in our data.

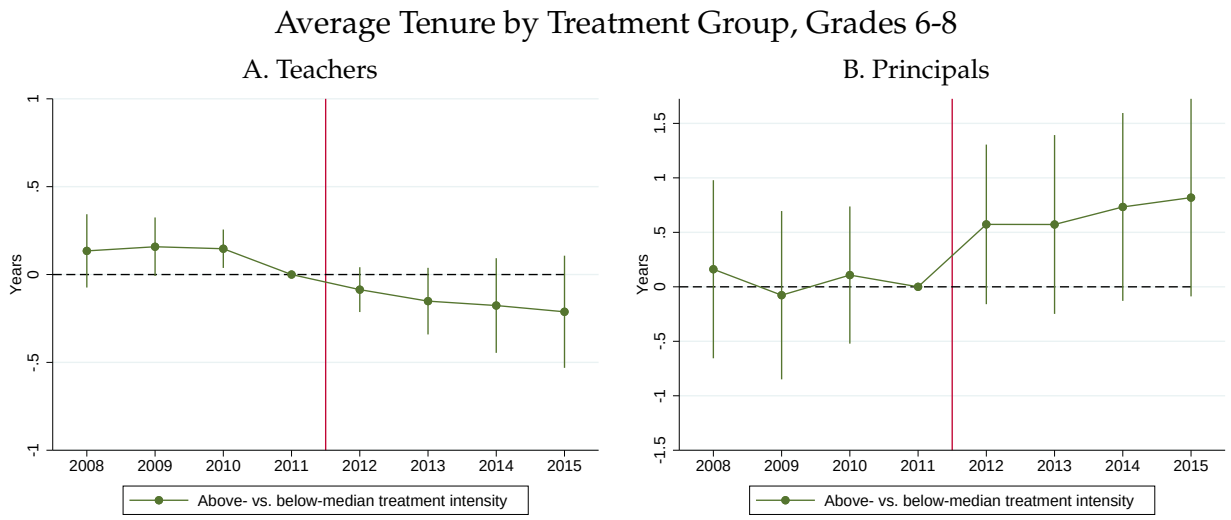

Figure 10. This figure shows the effects of the 2012 discipline reform on principal and teacher tenure. Panel A plots estimated treatment effects from Equation 3 for teacher tenure; Panel B plots the same for principal tenure. Each point measures the gap in tenure between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

6.1.1 Bounding Direct Effects with Sharp Timing

As a supplementary exercise, we estimate an upper bound for short-term direct effects in Appendix H. Our methodology compares two similar groups of students: those who were suspended just before each standardized test, and those suspended just afterward. Intuitively, the test scores of this latter group could not have been affected by their suspensions. We note that this approach cannot be used to measure the *long-term* effects of suspension, since all the students we compare are suspended before longer-term outcomes are realized. However, it would take very aggressive assumptions about long-term direct effects to produce the aggregate test score gains we observe.

The results imply that the short-term effect of each suspension is small, which in turn suggests that mechanical disruption from missing class while serving a suspension is not driving our results. We find that receiving a principal's suspension has no more than a 0.03 standard deviation causal impact on a student's math score later that year. The upper bound for the impact of superintendent's suspensions is larger, but even those longer punishments for more serious infractions have at most a 0.12 standard deviation impact. We obtain qualitatively similar results for reading.

6.2 Principal and Teacher Turnover

Principals and teachers are the primary drivers of variation in discipline policy and school culture within New York City, since all schools face the same set of district policies. In

conjunction with the change in the discipline code in 2012, it is conceivable that principals and teachers who favored strict, punitive discipline could have been replaced by others whose views were more aligned with the effort to reduce suspension rates.

Figure 10 shows the results when we estimate Equation 3 using principal and teacher tenure, which suggest otherwise. First, Panel A shows a slight relative decline in teacher tenure in the High Treatment group. This mirrors gradual declines in salary and experience (see Appendix G). However, these are very small changes, and there were no shifts in

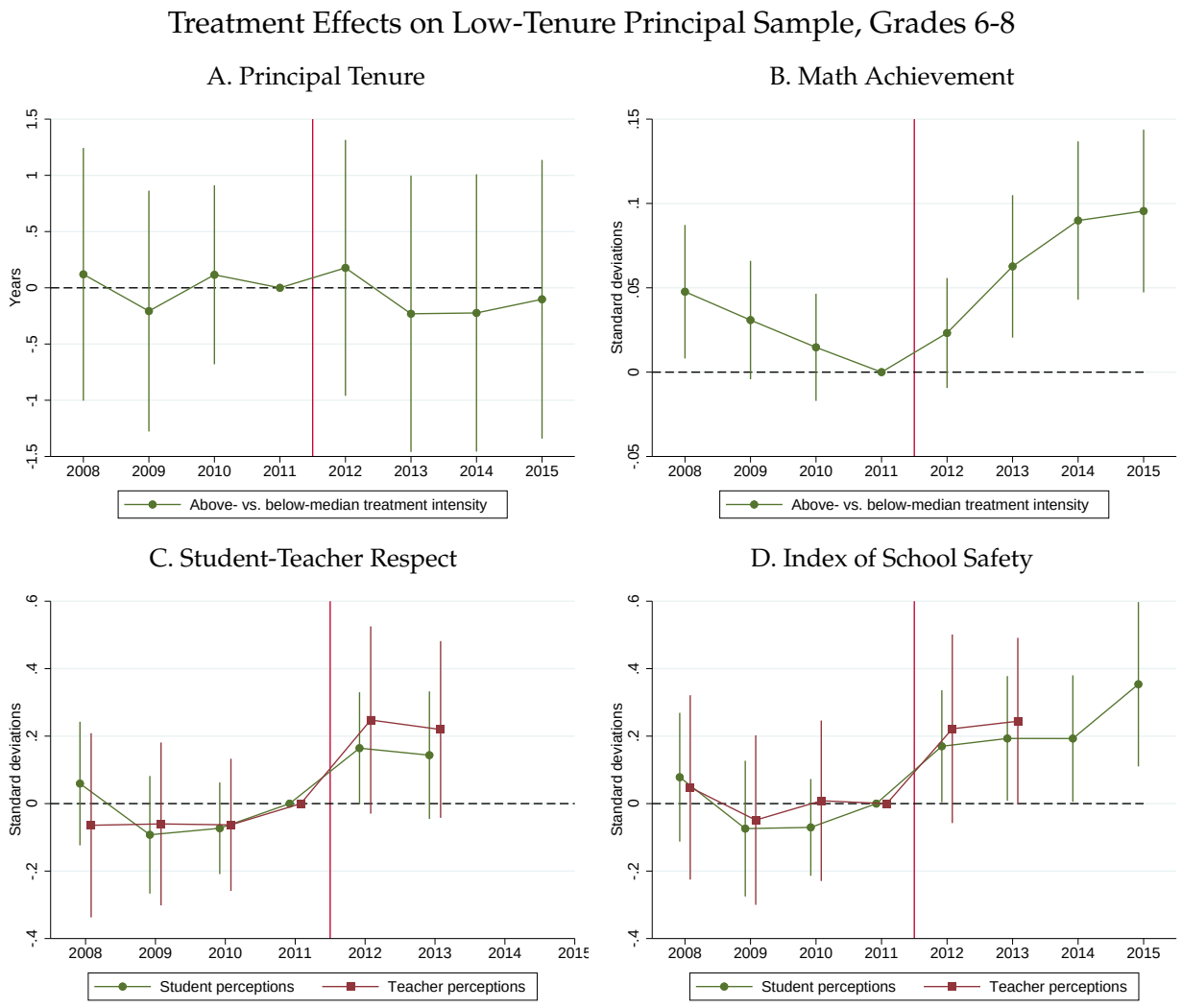

Figure 11. This figure shows the effects of the 2012 discipline reform on principal tenure, math scores, student-teacher respect, and perceptions of safety for a sub-sample of schools that excludes long-tenured principals (over 10 years). Each point measures the gap in the relevant outcome between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

the trend at the time of the reform.[34] Second, Panel B shows that principal tenure *increased* in High Treatment schools relative to Low Treatment schools, which is inconsistent with the replacement of high-suspension principals. This is driven by departures of a small number of principals with very long tenure from Low Treatment schools. In Figure 11, we assess the impact of these principals by dropping all schools with long-tenured principals (over 10 years) from both treatment groups. In this sample, there is no relative change in tenure, yet our treatment effects on test scores and school culture are even larger.

6.3 Student Turnover

Panel A of Figure 12 shows a similar analysis of impacts on enrollment levels. Before 2012, school-grades in the High Treatment group had been shrinking slightly relative to those in the Low Treatment group. They began to expand slightly after the reform, although the shift is small relative to the average size of a school-grade. To the extent that these pre-trends would have continued, this suggests a positive treatment effect on enrollment, which is consistent with students sorting into schools with improving cultures. Nonetheless, Panel B of the figure does not suggest changes in student composition based on prior test scores. Our results are therefore unlikely to have been driven by mechanical changes in composition, or by peer effects from any such compositional change.

6.4 School Resources

The 2012 reform encouraged a shift toward restorative interventions instead of suspensions. Such counseling-based interventions are resource-intensive, since they require investment in training and personnel. If additional resources were allocated to High Treatment schools, this could itself contribute to culture and test score improvements.

If anything, High Treatment school-grades became *worse* off in terms of per-student funding and the number students per counselor. Panel A of Figure 13 plots the coefficients from Equation 3 for total per-pupil expenditure. Funding decreased very slightly in the High Treatment group relative to the Low Treatment group. Panel B shows that there was little effect on the number of students per counselor. These results are consistent with complaints by teachers about the lack of new funding for counselors and psychologists to accompany the move toward non-punitive interventions (Baker, 2012).

[34] To the extent that any part of these changes are due to the reform, it could be the case that younger and relatively inexpensive teachers are more effective than the older teachers that they replace. However, the difference in efficacy would have to be very large to explain test score impacts of the size we find.

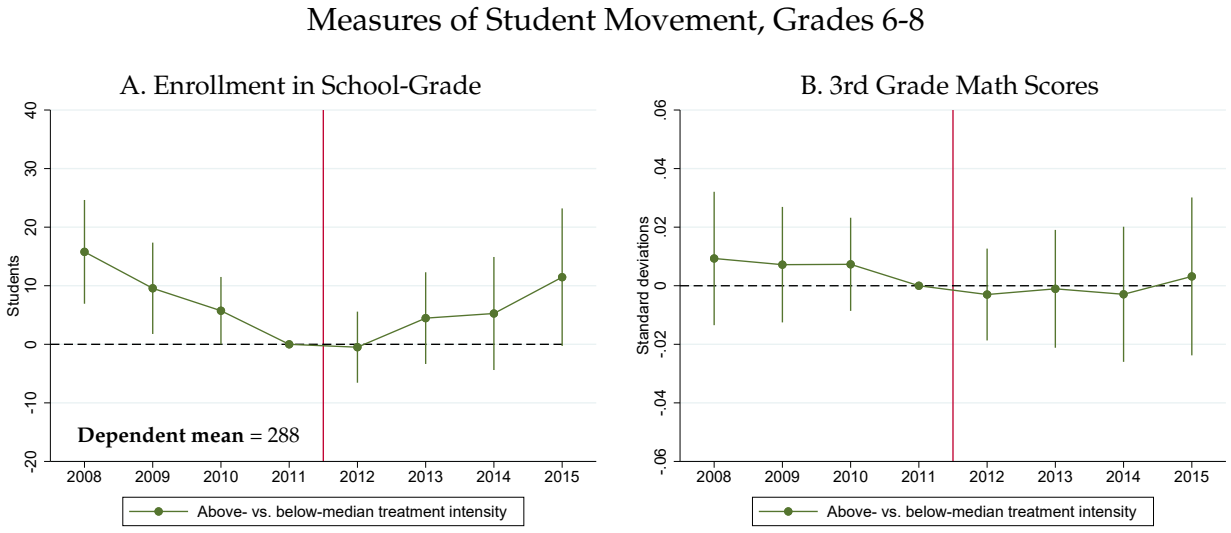

Figure 12. This figure shows the effects of the 2012 reform on enrollment and school-grade ability, as measured by math test scores from grade 3. Appendix Figure I18 shows results for reading. The green lines plot the estimated treatment effects ρ_k from Equation 3. Each point measures the difference between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

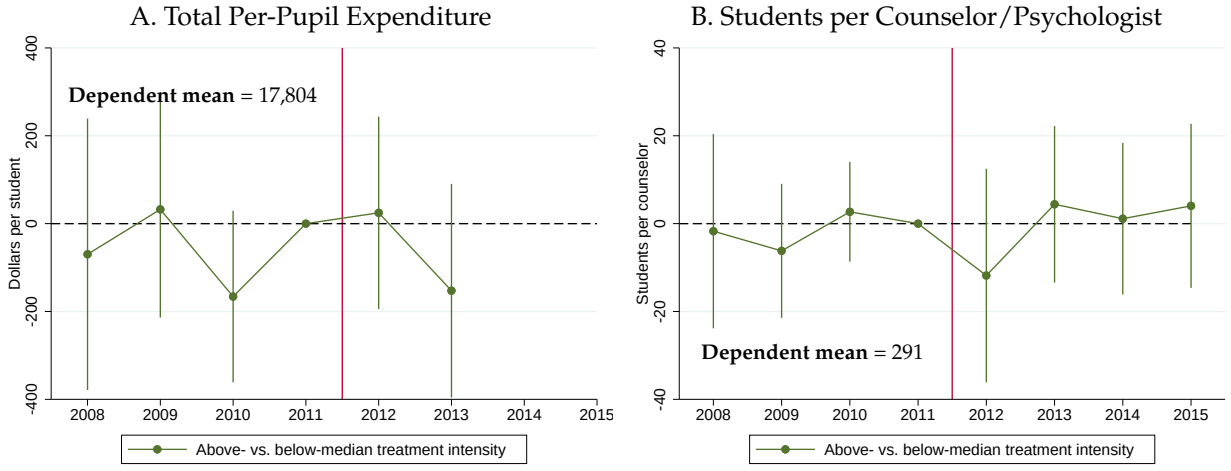

Figure 13. This figure shows the effects of the 2012 discipline reform on school resources. The green lines plot the estimated treatment effects ρ_k from Equation 3. Each point measures the difference between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

6.5 Correlates With Test Score Gains

To summarize the relationships between the test score gains and other variables, Table 6 shows correlations between changes in each variable and changes in test scores. In most cases, the period spans from 2011 to 2014. However, data availability limits us to 2013 for some variables. The results are as expected based on the analyses above.

The strongest correlates are our measures of culture and behavior: Increases in math achievement are positively correlated with improvements in survey measures of respect, safety, and behavior; and with reductions in the frequency or disruptive incidents in the VADIR administrative data. The only other factor that stands out is cohort size. As we explain above, this could reflect students sorting into schools with improving cultures, but any such movement is not selected on student ability. It is also important to note that there is very little correlation between cohort size and test scores in general.

Table 6. Correlations With Test Score Gains Relative To 2011

Culture & behavior	Year	Correlation	Other variables	Year	Correlation
Respect (Teachers)	2013	0.12***	Teacher Tenure	2014	-0.05
Respect (Students)	2013	0.15***	Teacher Salary	2014	-0.02
Safety (Teachers)	2014	0.13***	Teacher Experience	2014	-0.02
Safety (Students)	2014	0.11***	School Funding	2013	-0.02
Behavior (Teachers)	2013	0.20***	Students/Counselor	2014	0.01
Behavior (Students)	2013	0.19***	Cohort Size	2014	0.12***
Disruptive Incidents	2014	-0.10***			

Table notes. This table shows correlations at the school-grade level between math test score gains since 2011 and changes in other variables over the same period. The end period changes depending on data availability as indicated by "Year" in the table. The left panel shows measures of culture and behavior, while the right panel shows other variables. Correlations are weighted by school-grade size using analytic weights. Data are from the New York City Department of Education, and include students from grades 6 through 8.

7 Conclusion

Recent reforms across the United States have aimed to reduce suspension use. This paper provides evidence that a reform eliminating suspensions for disorderly behavior in New York City led to significant gains in test scores for students in schools that were more affected by the change, relative to other schools. Moreover, these benefits were obtained at minimal financial cost, and we found no evidence of a trade-off between academic achievement and safety or disruptive behavior. Our results suggest that the gains were driven by cultural changes that benefited a wide range of students, even those who would not have been suspended under the previous regime. By contrast, we rule out the possibility that any significant part of the gains came from the elimination of the direct impact

of suspension on students who would themselves have been suspended.

More broadly, this paper contributes to our understanding of the factors that make schools effective. It is well-documented that students' academic achievement can be improved by effective teachers (Chetty *et al.*, 2011), and by high-performing charter schools (Angrist *et al.*, 2013). We also have some appreciation for the package of practices that makes such schools effective; This tends to include strict discipline, along with frequent teacher feedback, data-driven instruction, high-dosage tutoring, and increased instructional time (Dobbie and Fryer, 2013; Fryer, 2014). However, our results suggest that instituting a strict discipline code by itself can be harmful to students, at least in the context of the New York City public schools that we study.

Our results will be encouraging for those who seek to reduce the reliance of schools on suspensions and other exclusionary punishments. The improvements we see in school culture contrast sharply with stated justifications behind strict discipline policies and high suspension rates. However, the details of implementation may matter. Our findings are consistent with evidence from Philadelphia (Lacoe and Steinberg, 2019) and Massachussets (Cleveland, 2022). They also align with work by Bacher-Hicks *et al.* (2019) which demonstrates a causal link between high suspension rate schools and negative criminal justice outcomes. But in contrast, Pope and Zuo (2023) argue that test scores fell in the Los Angeles Unified School District in response to efforts to reduce suspension rates over a decade. As they note, the difference in results may stem from the gradual and less-centralized nature of the L.A. reform, which featured significant local discretion.

Similarly, we suggest caution when generalizing from our results to other types of discipline reform, especially those that target higher-level suspensions. The 2012 reform in New York City that we study here was the first step in easing a very strict discipline code. The reform targeted the most discretionary suspensions, which were most likely to be perceived as overly harsh or unfair. By contrast, suspension may be necessary for more serious infractions, especially those that pose physical safety risks to other students. It is an open question whether changes to policies that target punishments for these infractions would lead to cultural and achievement gains similar to those we document here.

References

ADUKIA, A., FEIGENBERG, B. and MOMENT, F. (2023). From Retributive to Restorative: An Alternative Approach to Justice. *NBER Working Paper No. 31675*.

AMERICAN PSYCHOLOGICAL ASSOCIATION ZERO TOLERANCE TASK FORCE (2008). Are

Zero Tolerance Policies Effective in the Schools? An Evidentiary Review and Recommendations. *American Psychologist*, **63** (9), 852–862.

ANDERSON, K. P. (2018). Inequitable compliance: Implementation failure of a statewidestudent discipline reform. *Peabody Journal of Education*, **93**, 244–1263.

— and MCKENZIE, S. (2022). Local implementation of state-level discipline policy: Administrator perspectives and contextual factors associated with compliance. *AERA Open*, **8**, 1–20.

—, RITTER, G. W. and ZAMARRO, G. (2019). Understanding a vicious cycle: The relationship between student discipline and student academic outcomes. *Educational Researcher*, **48** (5), 251–262.

ANFARA JR., V. A., EVANS, K. R. and LESTER, J. N. (2013). Restorative justice in education: What we know so far. *Middle School Journal*, **44** (5), 57–63.

ANGRIST, J. D., PATHAK, P. A. and WALTERS, C. R. (2013). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*, **5** (4), 1–27.

AUGUSTINE, C. H., ENGBERG, J., GRIMM, G. E., LEE, E., WANG, E. L., CHRISTIANSON, K. and JOSEPH, A. A. (2018). Can restorative practices improve school climate and curb suspensions? an evaluation of the impact of restorative practices in a mid-sized urban school district. *RAND Research Report*.

BACHER-HICKS, A., BILLINGS, S. B. and DEMING, D. J. (2019). The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime.

BACKES, B., COWAN, J., GOLDHABER, D. and THEOBALD, R. (2022). Teachers and school climate: Effects on student outcomes and academic disparities. *CALDER Working Paper No. 274-1022*.

BAKER, A. (2012). New Code Aims to Ease Suspensions of Students. *New York Times*.

BAKER-SMITH, E. C. (2018). Suspensions suspended: Do changes to high school suspension policies change suspension rates? *Peabody Journal of Education*, **93**, 190–206.

BÉNABOU, R. and TIROLE, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, **70** (3), 489–520.

— and TIROLE, J. (2006). Incentives and prosocial behavior. *American Economic Review*, **96** (5), 1652–1678.

BROOKOVER, W. B., SCHWEITZER, J. H., SCHNEIDER, J. M., BEADY, C. H., FLOOD, P. K. and WISENBAKER, J. M. (1978). Elementary school social climate and school achievement. *American Educational Research Journal*, **15** (2), 301–318.

CALLAWAY, B., GOODMAN-BACON, A. and SANT'ANNA, P. H. (2021). Difference-in-differences with a continuous treatment. *Arxiv Working Paper No. 2107.02637*.

CARRELL, S. E., HOEKSTRA, M. and KUKA, E. (2018). The Long-Run Effects of Disruptive Peers. *American Economic Review*, **108** (11), 3377–3415.

— and HOEKSTRA, M. L. (2010). Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids. *American Economic Journal: Applied Economics*, **2** (1), 211–228.

CHETTY, R., FRIEDMAN, J. N. and ROCKOFF, J. E. (2011). The Long-Term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. *NBER Working Paper No. 17699*.

—, — and — (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, **104** (9), 2593–2632.

CLEVELAND, C. (2022). Rethinking discipline: The effects of school disciplinary reform laws on adult and student behavior. *Working Paper*.

COBB-CLARK, D. A., KASSENBOEHMER, S. C., LE, T., MCVICAR, D. and ZHANG, R. (2015). Is There an Educational Penalty for Being Suspended from School? *Education Economics*, **23** (4), 376–395.

CRAIGIE, T.-A. (2022). Do school suspension reforms work? evidence from rhode island. *Educational Evaluation and Policy Analysis*, **4**, 667–688.

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*.

DECKER, G. and SNYDER, S. (2015). Long-awaited Discipline Policy Changes Further Restrict Suspensions, Restraints. *Chalkbeat New York*.

DISARE, M. (2016). New york city students show up in droves to question school discipline policy. *Chalkbeat*.

DOBBIE, W. and FRYER, R. G. (2013). Getting beneath the Veil of Effective Schools: Evidence from New York City. *American Economic Journal: Applied Economics*, **5** (4), 28–60.

DULAY, S. and KARADAĞ, E. (2017). The effect of school climate on student achievement. In E. Karadağ (ed.), *The Factors Effecting Student Achievement*, Springer Nature, pp. 199–213.

EDEN, M. (2017). *School Discipline Reform and Disorder: Evidence from New York City Public Schools, 2012-16*. Tech. rep., Manhattan Institute.

ELLINGSEN, T. and JOHANNESSON, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, **98** (3), 990–1008.

FEDERAL COMMISSION ON SCHOOL SAFETY (2018). Final Report of the Federal Commission on School Safety.

FRYER, R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, **129** (3), 1355–1407.

— (2017). Chapter 2 - the production of human capital in developed countries: Evidence from 196 randomized field experiments. In A. V. Banerjee and E. Duflo (eds.), *Handbook of Economic Field Experiments*, *Handbook of Economic Field Experiments*, vol. 2, North-Holland, pp. 95 – 322.

GNEEZY, U. and RUSTICHINI, A. (2000). A fine is a price. *Journal of Legal Studies*, **29** (1), 1–17.

GREGORY, A., SKIBA, R. J. and NOGUERA, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher*, **39**, 59–68.

HANUSHEK, E. A. (2020). Education production functions. In S. Bradley and C. Green (eds.), *Economics of Education*, 2nd edn., London: Academic Press, pp. 161–170.

HASHIM, A. K., STRUNK, K. and DHALIWAL, T. K. (2022). Justice for all? suspension bans and restorative justice programs in the los angeles unified school district. *Educational Evaluation and Policy Analysis*, **93**, 174–189.

HINZE-PIFER, R. and SARTAIN, L. (2018). Rethinking universal suspension for severe student behavior. *Peabody Journal of Education*, **93** (2), 228–243.

KINSLER, J. (2013). School Discipline: A Source or Salve for the Racial Achievement Gap? *International Economic Review*, **54** (1), 355–383.

KRAFT, M. A., MARINELL, W. H. and YEE, D. S.-W. (2016). School organizational contexts, teacher turnover, and student achievement. *American Educational Research Journal*, **53** (5), 1411–1449.

KRUEGER, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, **114** (2), 497–532.

KUTSYURUBA, B., KLINGER, D. A. and HUSSAIN, A. (2015). Relationships among school climate, school safety, and student achievement and well-being: A review of the literature. *Review of Education*, **3** (2), 103–135.

LACOE, J. and STEINBERG, M. P. (2018). Rolling Back Zero Tolerance: The Effect of Discipline Policy Reform on Suspension Usage and Student Outcomes. *Peabody Journal of Education*, **93** (2), 207–227.

— and — (2019). Do Suspensions Affect Student Outcomes? *Educational Evaluation and Policy Analysis*, **41** (1), 34–62.

LAVY, V. and SCHLOSSER, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, **3** (2), 1–33.

LAZEAR, E. (2001). Educational production. *Quarterly Journal of Economics*, **116** (3), 777–803.

MILLER, J., OFER, U., ARTZ, A., BAHL, T., PHENIX, D., SHEEHAN, N. and THOMAS, H. A. (2011). *Education Interrupted: The Growing Use of Suspensions in New York City's Public Schools*. Tech. rep., New York Civil Liberties Union.

MORRIS, E. W. and PERRY, B. L. (2016). The Punishment Gap: School Suspension and Racial Disparities in Achievement. *Social Problems*, **63** (1), 68–86.

MORRISON, K. (2018). Students' perceptions of unfair discipline in school. *Journal of Classroom Interaction*, **53**, 21–45.

MUKHERJEE, E. (2007). *Criminalizing the Classroom: The Over-Policing of New York City Schools*. Tech. rep., New York Civil Liberties Union.

NEW YORK CITY DEPARTMENT OF EDUCATION (2004). Regulation of the Chancellor A-443.

NOLTEMEYER, A. L., WARD, R. M. and MCLOUGHLIN, C. (2015). Relationship Between School Suspension and Student Outcomes: A Meta-Analysis. *School Psychology Review*, **44** (2), 224–240.

PALLAS, A. M. (1988). School climate in american high schools. *Teachers College Record*, **89**, 541–554.

PAS, E. T., RYOO, J. H., MUSCI, R. J. and BRADSHAW, C. P. (2019). A state-wide quasi-experimental effectiveness study of the scale-up of school-wide positive behavioral interventions and supports. *Journal of School Psychology*, **73**, 41–55.

PERRY, B. L. and MORRIS, E. W. (2014). Suspending Progress: Collateral Consequences of Exclusionary Punishment in Public Schools. *American Sociological Review*, **79** (6), 1067–1087.

POPE, N. G. and ZUO, G. W. (2023). Suspending suspensions: The education production consequences of school suspension policies. *Economic Journal*, **133** (653).

SKIBA, R. J. and KNESTING, K. (2001). Zero tolerance, zero evidence: An analysis of school disciplinary practice. *New Directions for Youth Development*, **2001** (92), 17–43.

SORENSEN, L. C., BUSHWAY, S. D. and GIFFORD, E. J. (2022). Getting tough? the effects of discretionary principal discipline on student outcomes. *Education Finance and Policy*, **17** (2).

STEINBERG, M. P., ALLENSWORTH, E. and JOHNSON, D. W. (2011). Student and teacher safetyin chicago public schools: The roles of community context and school social organization. *University of Chicago Consortium on School Research*.

— and LACOE, J. (2017). What do we know about school discipline reform? assessing the alternatives to suspensions and expulsions. *Education Next*, **17** (1), 44–52.

WAY, S. M. (2011). School discipline and disruptive classroom behavior: The moderating effects of student perceptions. *Sociology Quarterly*, **52**, 346–375.

WELSH, R. O. (2023). Up the down escalator? examining a decade of school discipline reforms. *Children and Youth Services Review*, **150**, 1–13.

ZIMMERMAN, A. (2016). New york city school suspensions continue to plummet, but stark disparities persist. *Chalkbeat*.

A Decomposition of Covariation using Fixed Effects

In this appendix, we show that most of the negative covariation between suspensions and test scores is explained by differences between schools and between the types of students who do and do not get suspended within a school. Yet, student performance is still lower in years with a suspension relative to years without a suspension. Like Lacoe and Steinberg (2019), we exploit the panel structure of our data through regressions with individual student fixed effects and find similar results in New York as they did in Philadelphia. Comparing the results here to those in Appendix H, we show that our timing analysis offers meaningful improvements over fixed effects, especially for shorter suspensions.

Specifically, we estimate the following model on data prior to the removal of suspensions for Level 2 infractions in 2012:

$$y_{ijt} = \underbrace{\eta_i + \theta_j + \gamma_t}_{\textbf{Fixed effects}} + \beta \underbrace{\mathbb{1}(s_{ijt} > 0)}_{\textbf{Suspended}} + \epsilon_{ijt} \tag{4}$$

where y_{ijt} is the test score of student i in school-grade j in year t; $\mathbb{1}\left(s_{ijt} > 0\right)$ indicates that student i was suspended during year t; and η_i, θ_j, and γ_t are fixed effects.

The coefficient β measures the association between suspension rates and test scores, conditional on student characteristics. As such, this approach eliminates the possibility that β simply captures variation across students (e.g., due to differences in family background) that leads to both lower test scores and higher suspension rates. Instead, it is an estimate of the extent to which being suspended predicts *lower than usual* test scores for the affected student – holding fixed her grade and the school she attends.

Since longer suspensions for more severe behavior may have different effects, we also estimate versions with separate indicators for infraction severity or type of suspension:

$$y_{ijt} = \sum_{k=2}^{5} \overbrace{\beta_k \mathbb{1}(s_{ijt}^{Lk} > 0)}^{\textbf{Level } k \textbf{ suspension}} + \overbrace{\eta_i + \theta_j + \gamma_t}^{\textbf{Fixed effects}} + \epsilon_{ijt} \tag{5}$$

$$y_{ijt} = \underbrace{\beta_{\mathrm{P}} \mathbb{1}(s_{ijt}^{\mathrm{P}} > 0) + \beta_{\mathrm{S}} \mathbb{1}(s_{ijt}^{\mathrm{S}} > 0)}_{\textbf{Suspensions of each type}} + \underbrace{\eta_i + \theta_j + \gamma_t}_{\textbf{Fixed effects}} + \epsilon_{ijt} \tag{6}$$

where $\mathbb{1}(s_{ijt}^{L2} > 0)$ is an indicator for having at least one suspension for Level 2 behavior during that school year, and so on, and s_{ijt}^{P} and s_{ijt}^{S} refer to principal's and superintendent's suspensions. The coefficient β_x is the effect of having a particular level/type of suspension conditional on one's record of suspensions of other levels/types.

The results are shown in Tables A1 and A2 for math and reading. Column (1) in each table shows unconditional correlations between suspensions and test scores for middle school students. Students with at least one suspension in a given year perform 0.64σ worse on average in math and 0.51σ worse in reading. Adding school-grade and year fixed effects in Column (2) reduces the magnitudes of the effects by 20-25 percent. Most of the relationship between suspensions and test scores cannot therefore be explained by differences across schools. Rather, even within school-grades, students who get suspended fare much worse than those who do not get suspended.

Even when student fixed effects are added in Column (3), there remains a negative and statistically significant relationship between suspensions and test scores. The magnitudes of the coefficients are reduced, but effect sizes are meaningful.[35] Students who are suspended at least once score 0.06σ lower in math and 0.03σ lower in reading. These results are similar to Lacoe and Steinberg (2019), who found effects of 0.04σ in math and reading for middle school students in Philadelphia. In New York, these negative relationships are stronger for suspensions for higher-level infractions and for superintendent's suspensions relative to principal's suspensions, as shown in Columns (4) and (5).[36]

As with our timing analysis in Appendix H, these fixed effects estimates are upper bounds on the causal effect of a suspension: we do not observe student-specific shocks that may contribute to both misbehavior and poor exam performance, and these shocks are likely to operate in the same direction as the suspension effect. Comparing the math results in Figure H2 to Column (5) of Table A1, the fixed effects analysis *overstates* the direct effect of an individual suspension, although results are somewhat more precise. The implied upper bounds of 0.05σ and 0.11σ are meaningfully different from the timing analysis for principal's suspensions but not for superintendent's suspensions.

In summary, students within schools who get suspended are different from those who do not, yet a given student's performance is also lower in years in which they are suspended. Our timing analysis offers improvements over the fixed effects analysis in bounding effects on the suspended students, especially for shorter suspensions.

B Natural Experiment: Impact on Total Suspension Rate

Although the 2012 discipline reform explicitly banned suspensions for Level 2 infractions only, Panel B of Figure 4 shows that the drop in *total suspensions* was larger than the drop

[35] Similar estimates are produced by regressions with individual fixed effects but no school-grade effects.

[36] Lacoe and Steinberg (2019) found smaller and insignificant effects from suspensions for classroom disorder, which are broadly similar to our suspensions for Level 2 infractions.

Table A1. Regressions of Standardized Math Scores on Suspensions, Grades 6-8

	(1)	(2)	(3)	(4)	(5)
Any Suspension	-0.642	-0.513	-0.058		
	(0.012)	(0.009)	(0.004)		
Principal's				-0.043	
				(0.004)	
Superindendent's				-0.101	
				(0.007)	
Level 2					-0.025
					(0.011)
Level 3					-0.055
					(0.005)
Level 4					-0.049
					(0.004)
Level 5					-0.076
					(0.008)
School FE	No	Yes	Yes	Yes	Yes
Student FE	No	No	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Observations	1060457	1060457	881794	881794	881794
Clusters	1057	1057	1057	1057	1057
Adjusted R^2	0.026	0.241	0.792	0.792	0.792

Table notes. This table decomposes the unconditional correlation between math scores and suspensions using fixed effects. The dependent variable is the student's math score, standardized within a grade-year cell. Column (1) shows the unconditional correlation between math scores and suspensions for middle school students. Column (2) adds fixed effects for school-grade and year, and Column (3) adds individual student fixed effects to fully reflect β in Equation 4. Column (4) provides separate estimates of β for principal's and superintendent's suspensions using Equation 6. Column (5) provides separate estimates of β by infraction level using Equation 5. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Table A2. Regressions of Standardized Reading Scores on Suspensions, Grades 6-8

	(1)	(2)	(3)	(4)	(5)
Any Suspension	-0.509	-0.387	-0.028		
	(0.011)	(0.009)	(0.003)		
Principal's				-0.018	
				(0.003)	
Superindendent's				-0.050	
				(0.006)	
Level 2					-0.007
					(0.010)
Level 3					-0.025
					(0.005)
Level 4					-0.023
					(0.004)
Level 5					-0.035
					(0.008)
School FE	No	Yes	Yes	Yes	Yes
Student FE	No	No	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes
Observations	1038432	1038432	863192	863192	863192
Clusters	1057	1057	1057	1057	1057
Adjusted R^2	0.017	0.203	0.720	0.720	0.720

Table notes. This table decomposes the unconditional correlation between reading scores and suspensions using fixed effects. The dependent variable is the student's reading score, standardized within a grade-year cell. Column (1) shows the unconditional correlation between reading scores and suspensions for middle school students. Column (2) adds fixed effects for school-grade and year, and Column (3) adds individual student fixed effects to fully reflect β in Equation 4. Column (4) provides separate estimates of β for principal's and superintendent's suspensions using Equation 6. Column (5) provides separate estimates of β by infraction level using Equation 5. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

in Level 2 suspensions alone. Figure B1 shows the effects of the reform on other principal's suspensions and superintendent's suspensions. There is a clear drop in other principal's suspensions in the High Treatment group in 2012, separate from the steady convergence between the two groups over time. Policy effects on superintendent's suspensions are more muted, which is unsurprising given that superintendent's suspensions are rarer and typically less discretionary because they are for more serious misbehavior.

Effects on higher-level suspensions could come through either improvements in behavior or from changes in enforcement. The former are difficult to measure precisely in our setting, because available data are self-reported by schools and under-reporting could be important. Nonetheless, we provide evidence in Panel B of Figure 8 and in Figure 9 that both perceptions of behavior and actual behavior improve.

C Natural Experiment: Relative to All of New York State

Although data from the NYCDOE only include test scores for students in New York City, the same tests are also administered to students outside of the city but in New York State. In this appendix, we show how test scores evolved relative to students outside of the city, who were not affected by the discipline reform we study. To do so, we re-standardize the test scores to be relative to all of New York State using public information on the average and standard deviation of performance on these tests in each grade and year.

Figure C1 shows the results. For both reading and math, there are secular trends in the city average score relative to the state average. For math, the pre-reform trend is approximately linear. But at the time of the reform we see a large improvement in math scores for the Above-median treatment intensity group relative to the state average, with a smaller improvement for the Below-median group. To the extent that the trend in the city average compared to the state would have continued linearly, the movements in Figure C1 provide a measure of the absolute impact of the reform, rather than only the impact on the Above-median group relative to the Below-median group. The patterns are less clear for reading because the secular trends are not linear in the first place.

D Natural Experiment: Heterogeneity by Demographics

In this appendix, we estimate effects of the reform for particular subgroups of students who are more or less at risk of suspension on average. Panel A of Figure D1 shows that treatment effects on math scores for boys and girls – obtained by estimating Equation 1 for each group – are nearly identical. The results for reading are similar (see Appendix Figure

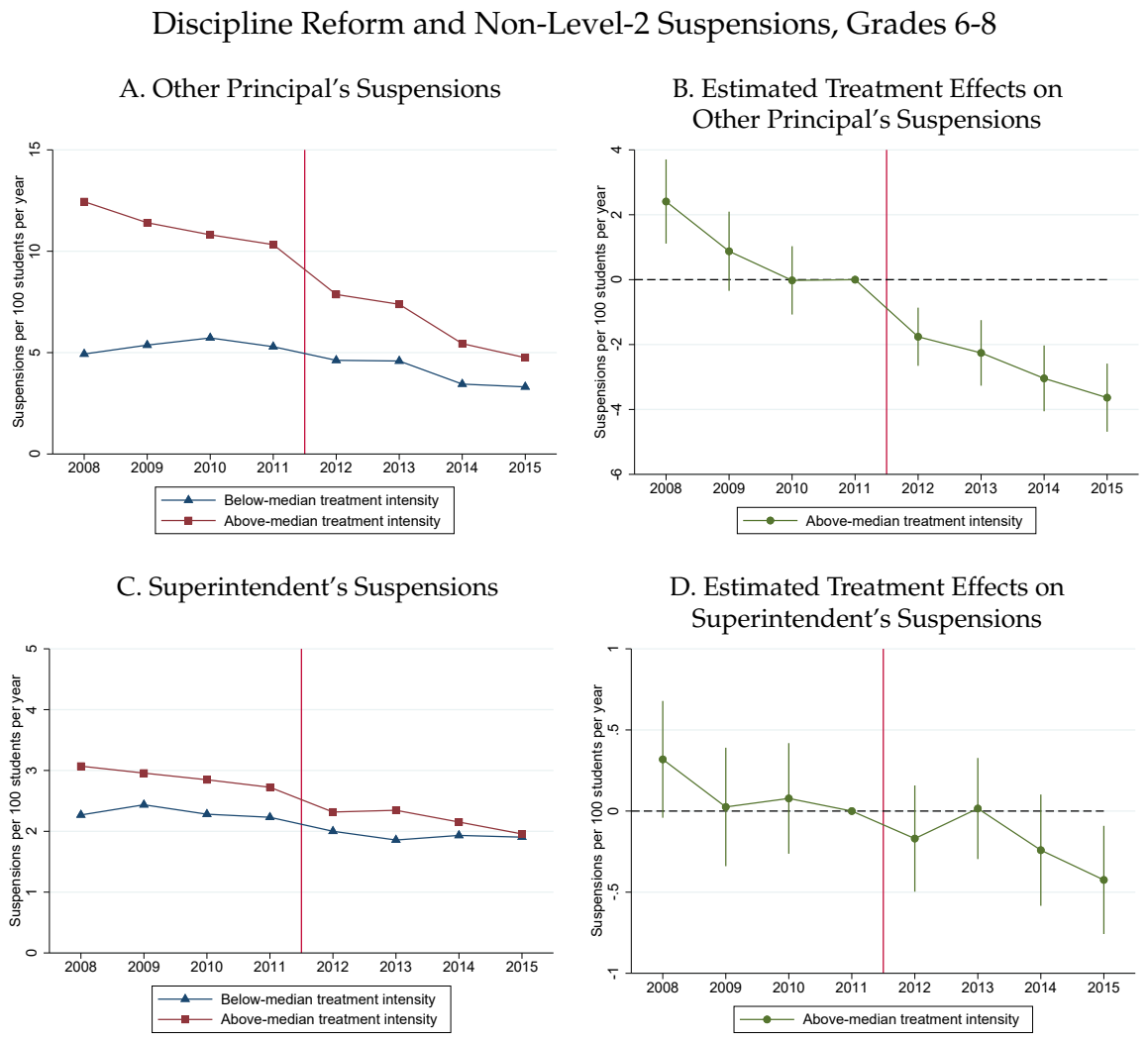

Figure B1. This figure shows the effects of the 2012 discipline reform on non-Level-2 suspensions. Panel A plots average rates of principal's suspensions for Level 3-5 infractions in each treatment group over time. Panel B plots the estimated treatment effects, ρ_k, from Equation 1. Each point measures the gap in suspension rates between the High Treatment and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. Panels C and D repeat the same analysis for superintendent's suspensions. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

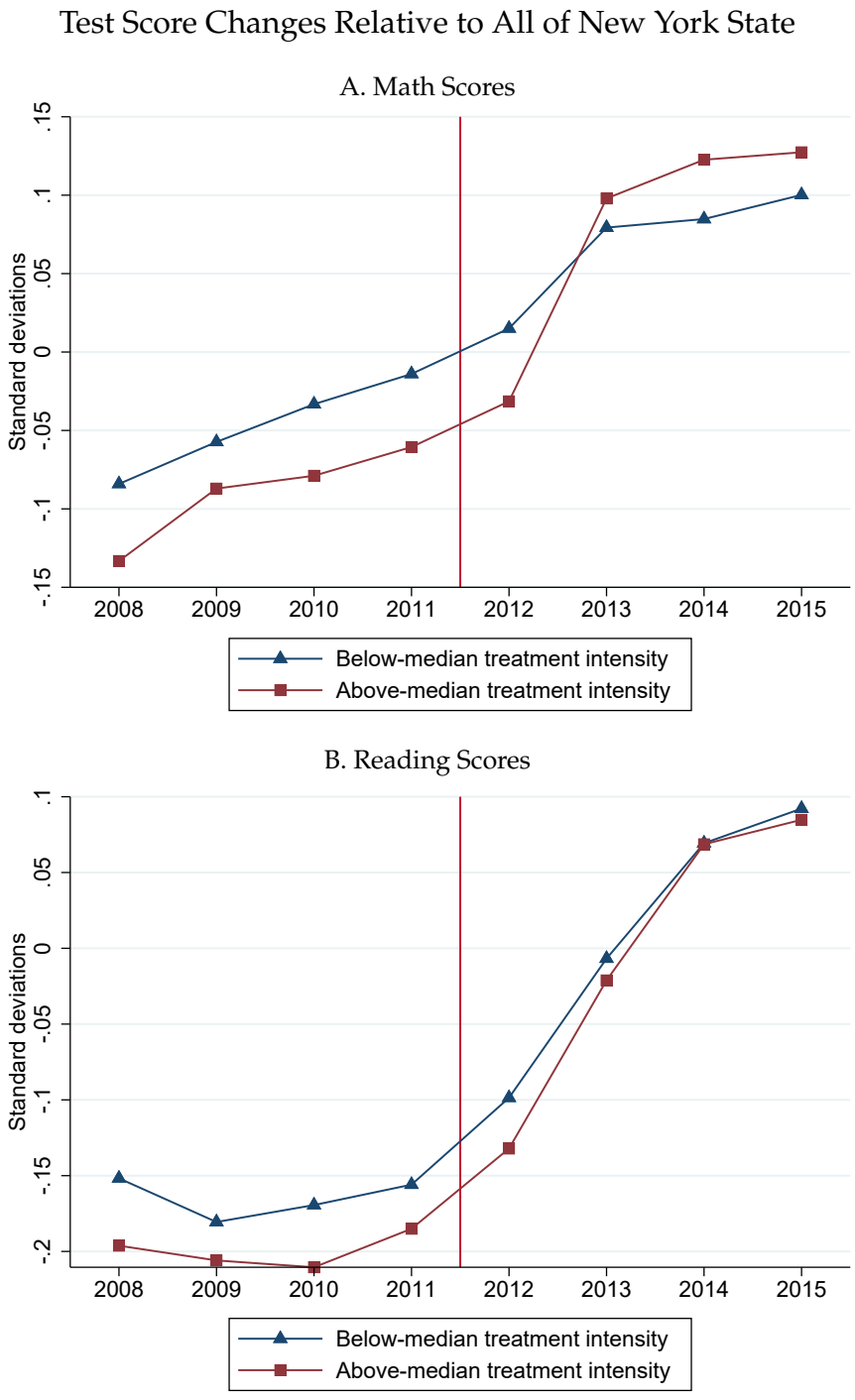

Figure C1. This figure shows the effects of the 2012 reform on math achievement when test scores are re-standardized to all of New York State rather than only New York City. Panel A plots average standardized math scores in each treatment group over time. Panel B plots reading scores. Test scores are standardized within the sample in grade-year cells. Data are from the New York City Department of Education, and include students from grades 6 through 8.

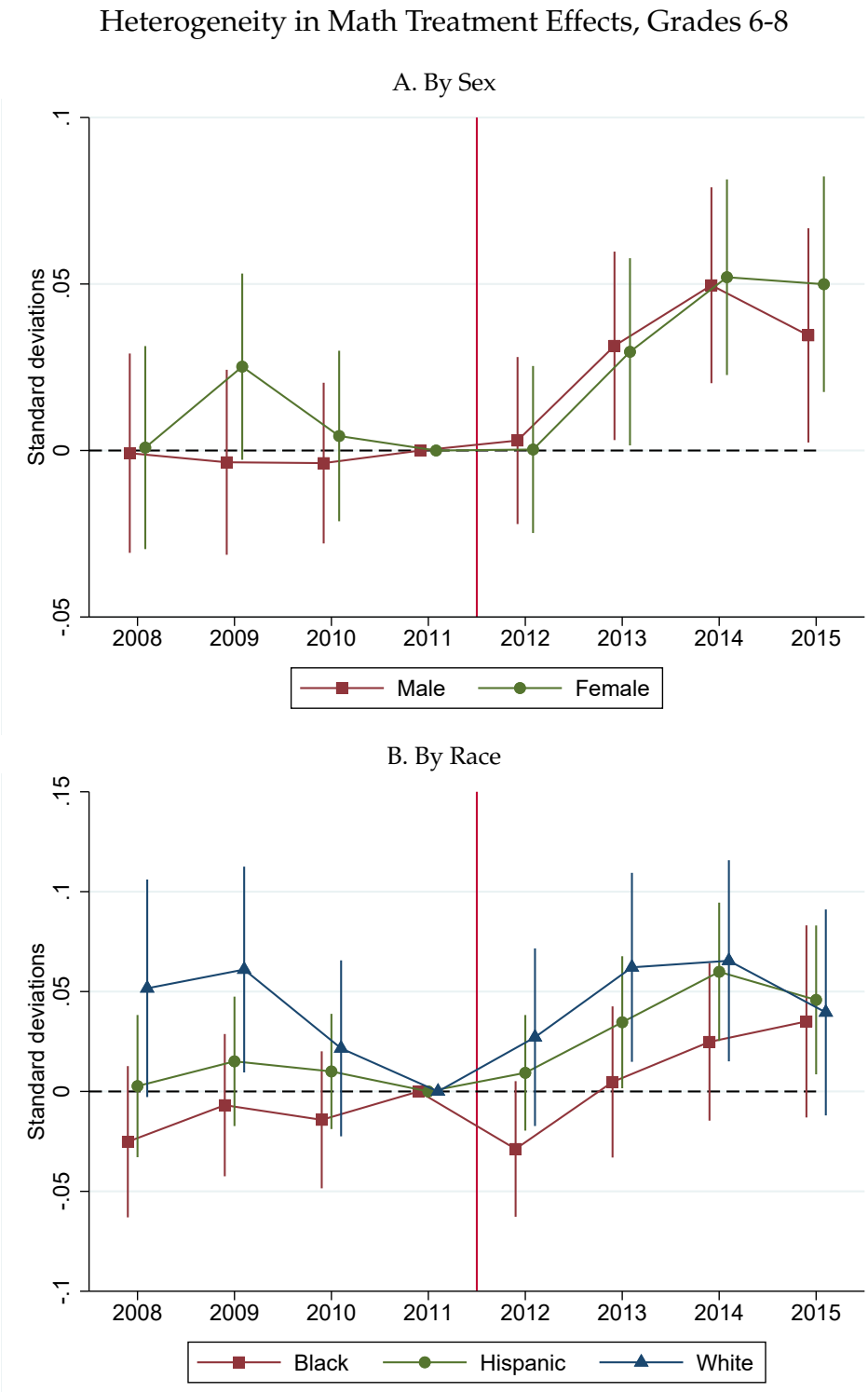

Figure D1. This figure shows the effects of the reform on math achievement for demographic subgroups. Panel A plots the estimated treatment effects ρ_k from Equation 1 for boys and girls. Panel B splits by racial group. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and other demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within grade-year cells. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6 through 8.

I6). This is despite boys being suspended twice as often as girls prior to the reform: from 2008 to 2011, middle school boys in New York City schools averaged 15.9 suspensions per 100 students for any infraction and 0.8 suspensions per 100 students for disorderly behavior, compared to 7.1 and 0.37 for girls.

The treatment effects by race in Panel B are also suggestive of smaller gains for black students, even though black students are much more likely to be suspended.[37] However, it is difficult to draw strong conclusions, since the scores of students in the High Treatment group and their same-race peers in the Low Treatment group do not move in parallel prior to the reform. The pre-trends are more consistent for reading (see Appendix Figure I6), with the results suggesting that the largest gains were for white students.

E Natural Experiment: Robustness Checks

Our estimated aggregate treatment effects in Section 4 are robust to a wide variety of alternative specifications, which we summarize in this Appendix.

E.1 Re-Balancing the Treatment Groups

As discussed in Section 4 and shown in Table 2, our High and Low Treatment groups are not perfectly balanced on demographics. This is not necessarily a problem for identification, but we show here that our estimates are nearly identical when we re-balance the Low Treatment group to have the same demographic mix as the High Treatment group. This suggests that our results cannot be driven by separate policies designed, for example, to benefit minority students relative to white students.

For each student i, we estimate the probability of being in a High Treatment school-grade ($T_i = 1$) with a simple logit regression:

$$p_i = \Pr(T_i = 1|X_i) = \frac{1}{1+e^{-\phi}}$$
$$\phi = \alpha + \beta X_i + \epsilon_i$$

where X_i includes individual-level covariates. We estimate α and β on data for 2008-2011 only, and then use these estimates to generate propensity scores $\hat{p}_i$ for 2008-2015.

We then translate our propensity scores into regression weights, w_i:

$$w_i = T_i + (1 - T_i)\frac{\hat{p}_i}{1-\hat{p}_i}$$

[37] The suspension rate for black students was 18.9 per 100 (1.0 per 100 for Level 2 infractions), compared to 11.6 per 100 (0.6 per 100) for Hispanic students, and 7.2 per 100 (0.4 per 100) for white students.

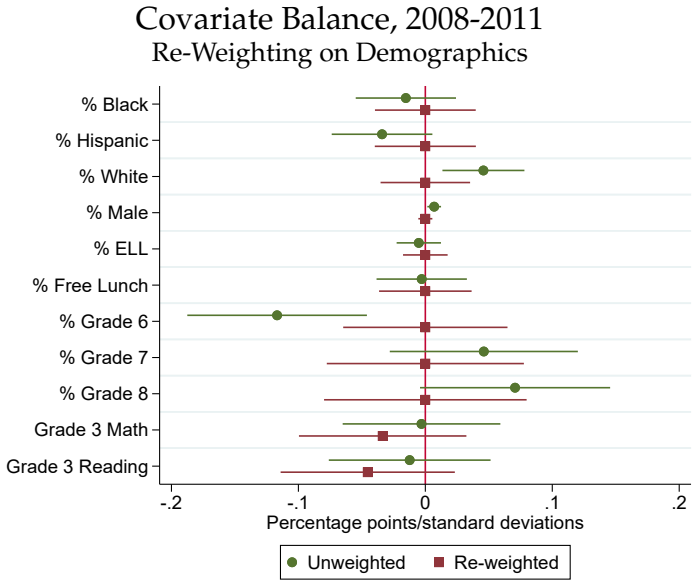

Figure E1. This figure plots the average pre-period difference between the High Treatment group and the Low Treatment group over several demographic characteristics. Positive point estimates reflect higher shares/scores in the High Treatment group. Average differences from our baseline, unweighted sample are in green, and average differences from the sample re-weighted along demographic characteristics are in red. Data are from the New York City Department of Education, and include students from grades 6 through 8.

where $\hat{p}_i$ is the propensity score for student i. For students in the High Treatment group, $w_i = 1$ by definition. But w_i re-balances the covariate distribution in the Low Treatment group to match that of the High Treatment group on average over 2008-2011.

We first re-balance on demographic characteristics - race, sex, ELL status, free lunch status, and grade. Figure E1 plots the difference beween the High and Low Treatment group means for the unweighted sample in green and the re-weighted sample in red. Because the demographic variables are discrete, we are able to achieve perfect balance on these measures over the pre-reform period. Figure E2 shows treatment effects for the unweighted and re-weighted samples. The estimates for the re-weighted sample in red are indistinguishable from our unweighted estimates in green, which tells us that differences in test score trends by demographic group are not driving our primary results.

Since balancing on demographics makes the sample less balanced on grade 3 test scores, Figure E3 repeats the same analysis, but re-balancing on grade 3 test scores rather than demographics. The re-weighted groups are then slightly less balanced on demographics. Regardless, Figure E4 shows that treatment effects are barely affected by the weighting. In fact, re-balancing on any combination of observables, including suspensions for non-Level-2 infractions, produces similar results.

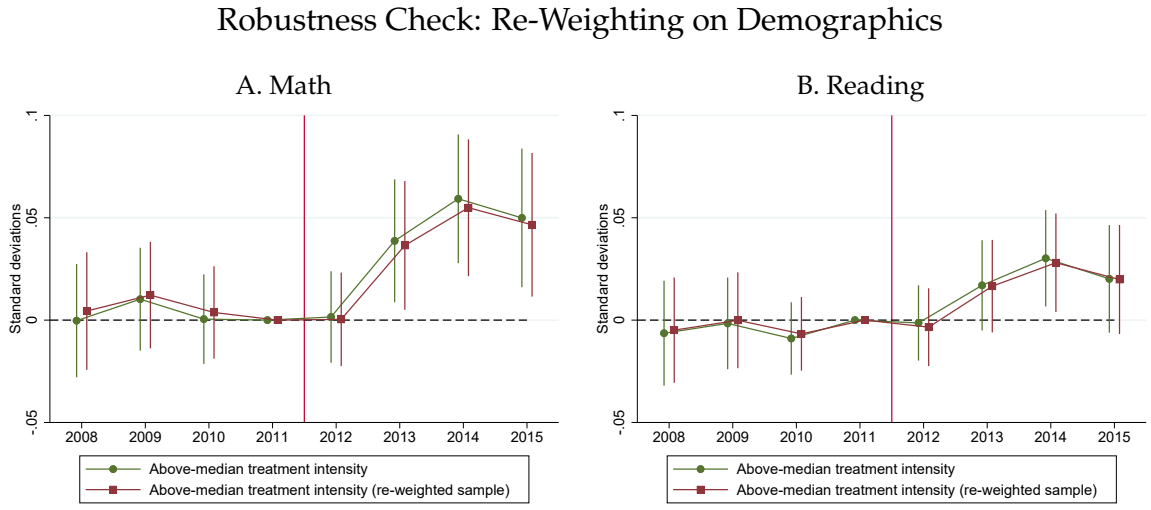

Figure E2. This figure compares estimated treatment effects on achievement in the baseline sample with a sample that was re-weighted to be balanced on demographics in the pre-period. In each panel, we plot two sets of treatment effects ρ_k, both estimated using Equation 1. Treatment effects using the baseline sample are in green and treatment effects using the re-weighted sample are in red. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the baseline sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

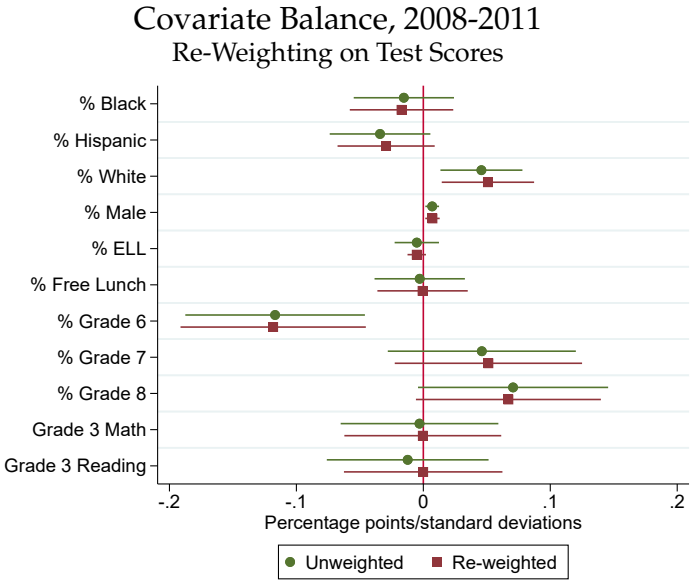

Figure E3. This figure plots the average pre-period difference between the High Treatment group and the Low Treatment group over several demographic characteristics. Positive point estimates reflect higher shares/scores in the High Treatment group. Average differences from our baseline, unweighted sample are in green, and average differences from the sample re-weighted by grade 3 test scores are in red. Data are from the New York City Department of Education, and include students from grades 6 through 8.

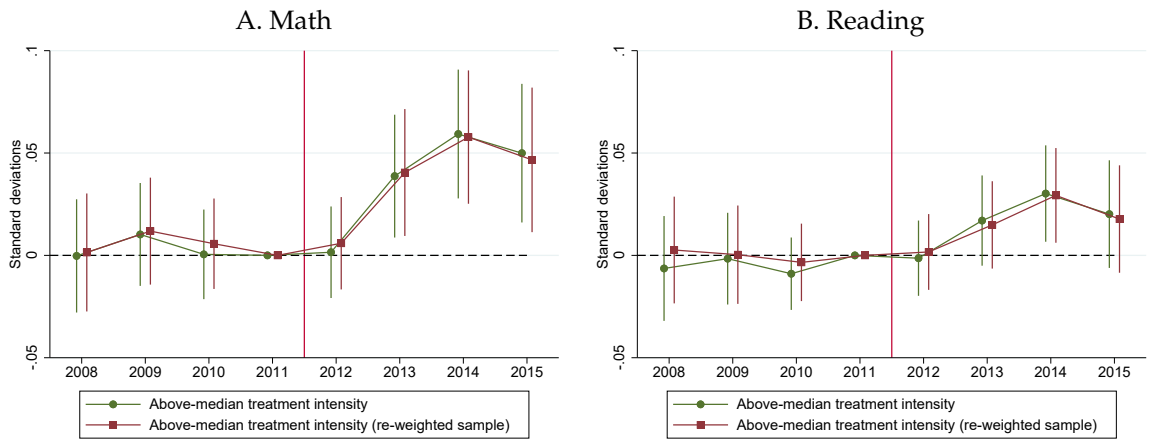

Figure E4. This figure compares estimated treatment effects on achievement in the baseline sample with a sample that was re-weighted to be balanced on grade 3 test scores in the pre-period. In each panel, we plot two sets of treatment effects ρ_k, both estimated using Equation 1. Treatment effects using the baseline sample are in green and treatment effects using the re-weighted sample are in red. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the baseline sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

E.2 Continuous Definition of Treatment

As an additional robustness check, we include results using a continuous definition of treatment. Specifically, we estimate the following specification:

$$y_{ijt} = \underbrace{\alpha_j + \gamma_t}_{\textbf{Fixed effects}} + \sum_{k \neq 2011} \rho_k \Big[\underbrace{\mathbb{1}(t=k)}_{\textbf{Time}} \times \underbrace{\hat{s}_j^{L2}}_{\textbf{Treatment}} \Big] + \underbrace{\beta X_{ijt}}_{\textbf{Controls}} + \varepsilon_{ijt} \tag{7}$$

where $\hat{s}_j^{L2}$ is the average actual suspension rate for Level 2 infractions in 2006 and 2007.

We prefer the discrete specification in Section 4 because we have no ex ante reason to think that treatment effects should be linear in the policy-induced reduction in suspensions, or in the pre-period suspension rates that we use as a proxy for treatment intensity. Nonetheless, Panel A of Figure E5 shows that our results for math are consistent with the discrete specification. The reading results, shown in Panel B, are less robust: we see a similar pattern of improvements from 2012-2015, but high pre-period suspension rates are associated with *lower* test scores in 2012 relative to 2011. However, the results are noisier for reading, and none of coefficients are significantly different from zero.

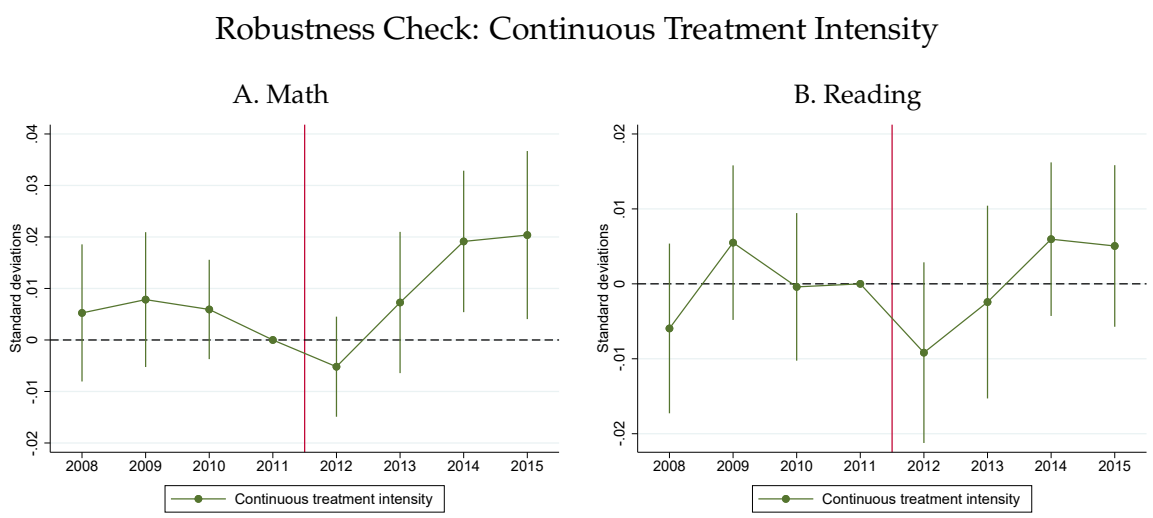

Figure E5. This figure shows the effects of the 2012 discipline reform on achievement when treatment intensity is measured continuously. Each panel plots the estimated treatment effects ρ_k from Equation 7. Each point measures the change in the slope of the test score-treatment intensity gradient relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

E.3 Alternative Prediction Period

We also include results using suspension rates over the entire 2006–2011 period to define our treatment groups. This may allow us to better distinguish between school-grades with similar suspension rates. However, the disadvantage of this longer base period is that there may be residual mean reversion at the time of treatment. When treatment is defined based on the shorter time period that we use in our main analysis, such mean reversion is not an issue. The results are shown in Figure E6, and are qualitatively unchanged.

E.4 Additional Treatment Group Bins

Conditional on using a discrete specification and estimating treatment intensity using 2006 and 2007 data, our choice of splitting at the median is unimportant. As one way of demonstrating this, Figure E7 provides an additional analysis in which we define treatment differently. First, we isolate the set of schools with zero suspension rates in 2006-07. Next, we replicate our main analysis for two different treatment groups compared to that reference group: school-grades with above-median and below-median suspension rates among those with any suspensions. For math, treatment intensity and math score improvements are tightly linked: the school-grades that improve between 2011 and 2014 are the ones with the highest treatment intensities. For reading, the relationship is positive but noisier. This is consistent with the two-group and continuous specifications.

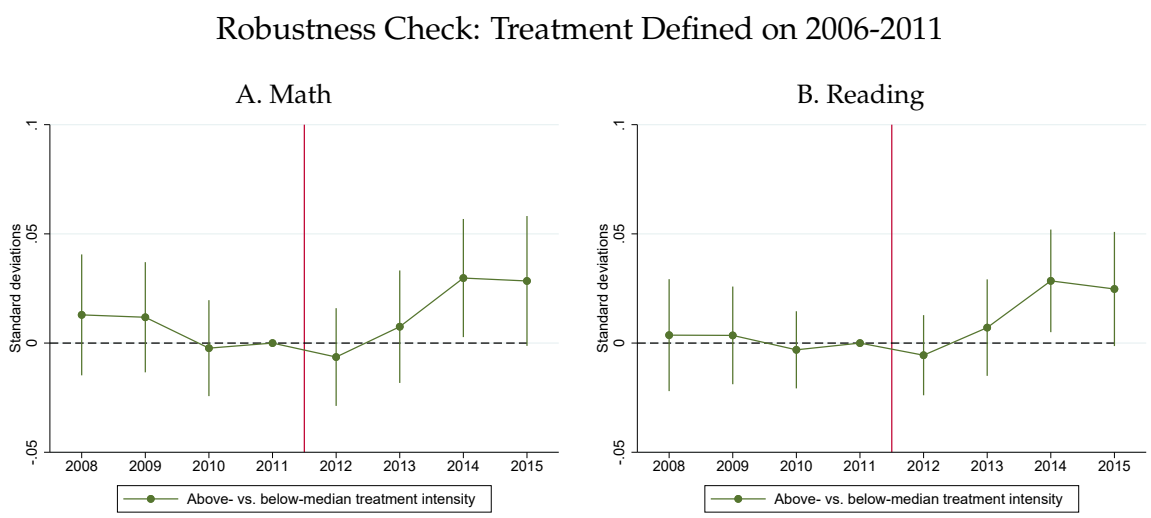

Figure E6. This figure shows the effects of the 2012 discipline reform on achievement when treatment intensity is estimated from average suspension rates for Level 2 infractions in 2006-2011. Each panel plots the estimated treatment effects ρ_k from Equation 1. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Figure E8 shows the results of a similar analysis but with four groups. Specifically, we again isolate schools with zero suspension rates in 2006-07. But we then split the remaining school-grades into three groups rather than the two in Figure E7. We then compare the most-treated group to the group with zero pre-period suspensions. The results for this most-treated group are quite similar to those for the Above-median treatment intensity group in Figure E7.

E.5 Treatment Based on All Suspensions

In our primary specification, we define treatment using suspension rates for Level 2 infractions. We view this as the logical choice, since these are the suspensions explicitly eliminated by the policy change. Moreover, higher-level suspensions fluctuate more, are more strongly correlated with test scores, and may also be more correlated with factors other than treatment intensity. This is reflected in less balanced High and Low Treatment groups when we define treatment based on the overall suspension rate, and in the fact that overall suspension rates were trending downward in the High Treatment group relative to the Low Treatment group before 2011.

Nonetheless, Figure E9 shows that the results are similar when treatment is defined using 2006-2007 suspension rates for all infractions. The test score gains are more muted, but the pre-trends are flat, and the dynamic pattern mirrors the main analysis.

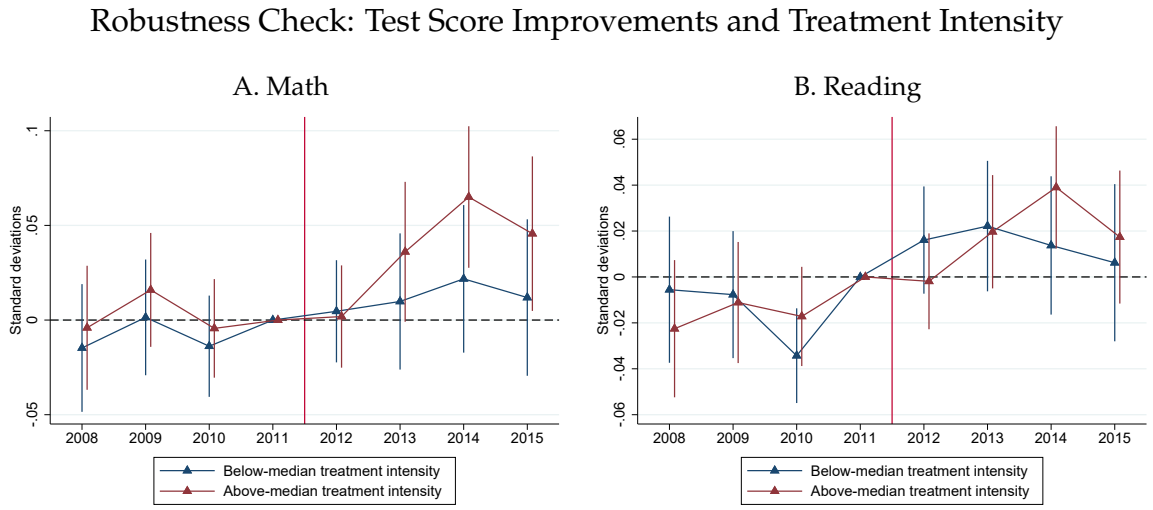

Figure E7. This figure shows the effects of the 2012 discipline reform on achievement when school-grades are divided into three groups. The reference group is the set of school-grades with zero pre-period suspensions. Those with positive pre-period suspension rates are then divided into those for which the rate is above versus below the median. Each point measures the gap in test scores between the relevant treatment group (Below-median or Above-median) and the zero suspension rate group relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

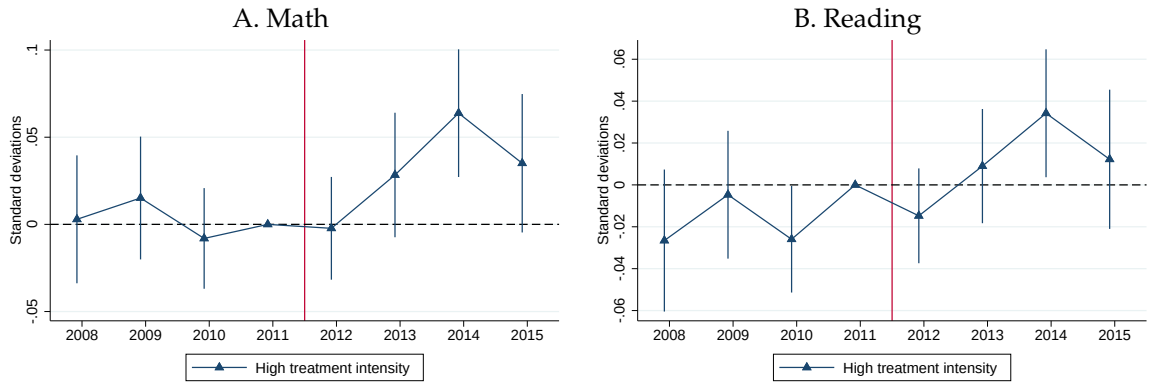

Figure E8. This figure shows the effects of the 2012 discipline reform on achievement when school-grades are divided into four groups. The reference group is the set of school-grades with zero pre-period suspensions. Those with positive pre-period suspension rates are then divided into terciles based on historical suspension rates. Each point in this graph measures the gap in test scores between the group with the highest historical suspension rate and the zero suspension rate group, relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. 2012 refers to the 2012-13 school year, and so forth. Data are from the New York City Department of Education, and include students from grades 6-8.

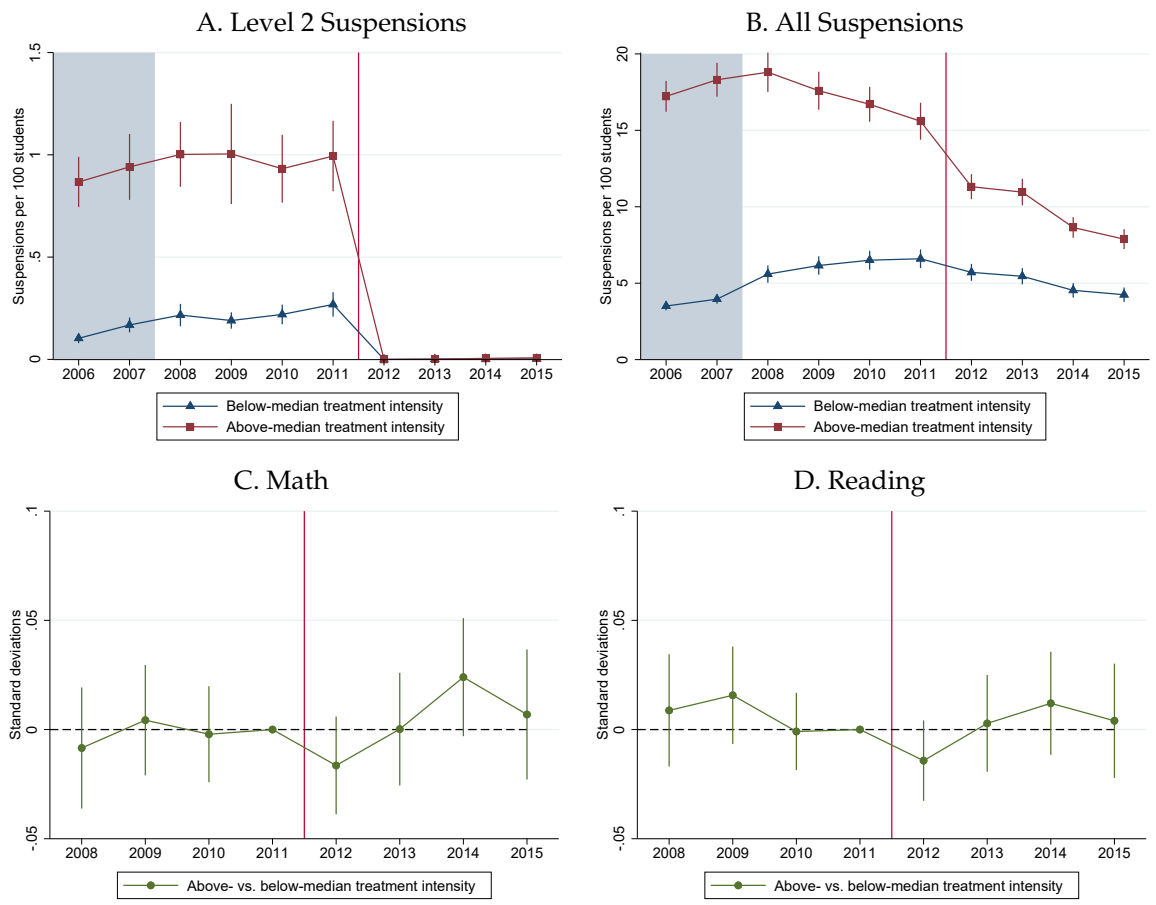

Figure E9. This figure shows the effects of the 2012 discipline reform on achievement when treatment intensity is estimated using suspension rates for *all* infractions in 2006-2007 rather than Level 2 infractions. Panels A and B plot average suspension rates for both Level 2 infractions (Panel A) and all infractions (Panel B) in the new treatment groups. Panels C and D plot the estimated treatment effects ρ_k from Equation 1. Each point measures the gap in test scores between the High Treatment and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

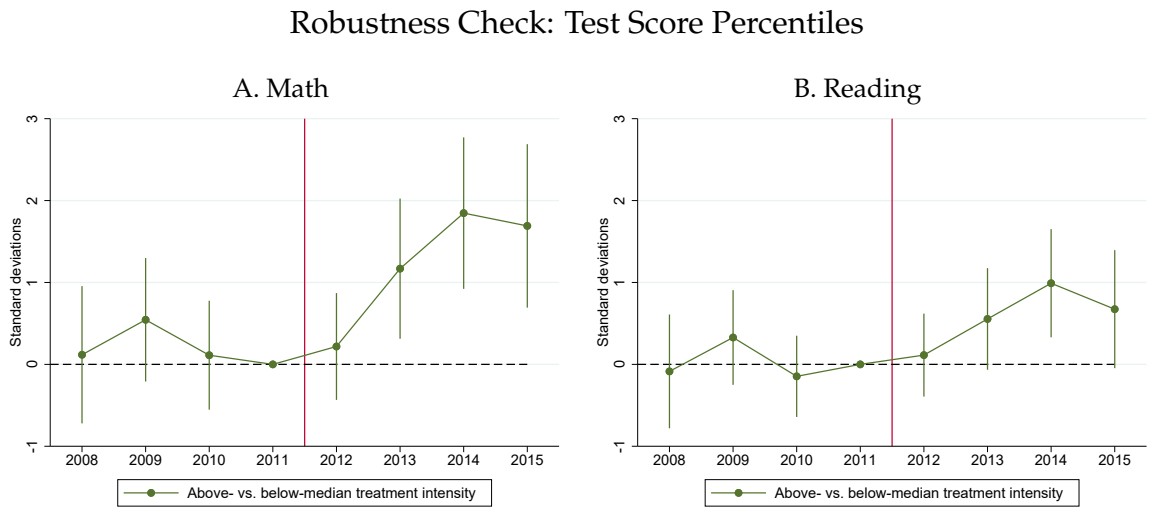

Figure E10. This figure shows the effects of the 2012 discipline reform on student math and reading achievement measured in percentiles rather than standardized scores. Each panel plots the estimated treatment effects ρ_k from Equation 1. Each point measures the gap in test scores between the High Treatment and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

E.6 Percentiles of Achievement

The patterns of treatment effects for both math and reading are unchanged when we use percentiles instead of standardized scores, as shown in Figure E10.

E.7 All School-Grades

Figure E11 shows that our results are robust to expanding our sample to include school-grades that enter or leave the data during the 2006-2015 period.

E.8 Adjustment for 2013 Testing Waiver

During the 2013 school year, the NYCDOE obtained a waiver from the federal government to allow accelerated students in grade 8 who were sitting the New York City Regents exam in mathematics to avoid "double-testing" by skipping the statewide grade 8 math exam. This produces a sharp reduction in the number of eighth grade students in our data in 2013. To avoid any concern that this sudden change in composition is driving our results, we adjust for it here by omitting any student who is eventually observed to sit the Regents exam for math in grade 8.[38] Excluding these students eliminates the jump in our

[38]This also requires us to drop 6th graders in 2014 and 2015 and 7th graders in 2015, because we do not observe whether they take the Regents exam. Results are again unchanged if we instead predict which of

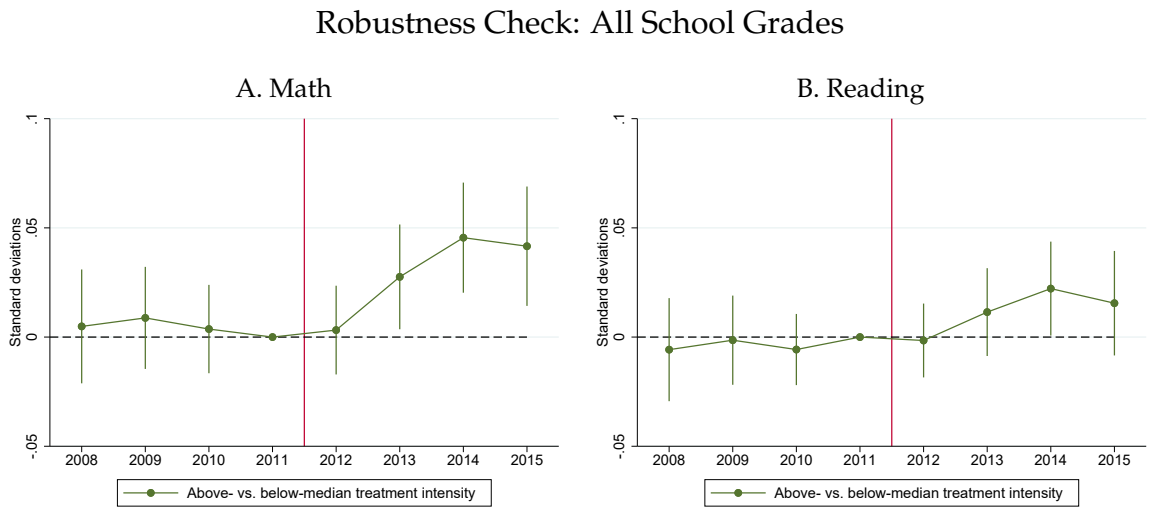

Figure E11. This figure shows the effects of the 2012 discipline reform on student math and reading achievement in a sample of *all* school-grades, not just those that form a balanced panel. Each panel plots the estimated treatment effects ρ_k from Equation 1. Each point measures the gap in test scores between the High Treatment and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and student demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

sample size in 2013. Moreover, since there is no sharp change in the percentage of students sitting the Regents exam in grade 8 (see Figure E12), the restriction effectively mitigates the impact of the change in composition in 2013. Treatment effects in the restricted sample are shown in Figure E13. They are qualitatively similar to our results in the full sample.

F Contemporaneous Policy Changes

We next discuss why our results are unlikely to be driven by the switch to Common Core in 2012 or the leadership transition to Mayor Bill de Blasio and Chancellor Carmen Fariña.

Common Core

Starting with the 2012 school year, all schools in New York State implemented a new set of learning standards called the Common Core, aimed at improving college and career readiness. This included revisions to standardized math and reading exams to better align with the new standards. Since standardized exams are a noisy measure of achievement, changing the standardized exam to a different noisy signal could produce a mechanical shift in the distribution of test scores. If this affected our treatment groups differently, it could in principle generate spurious treatment effects.

these students will take the Regents exam using their past math scores.

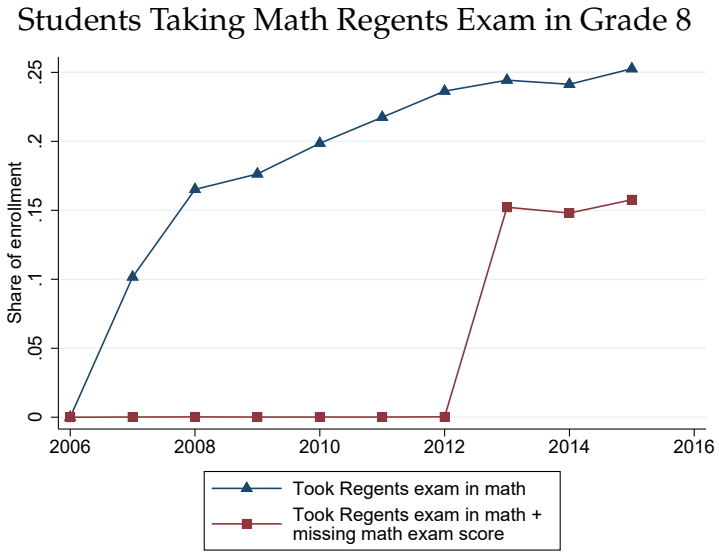

Figure E12. This figure shows the share of grade 8 students who take the Regents exam in mathematics, with and without taking the regular grade 8 exam. Data are from the New York City Department of Education.

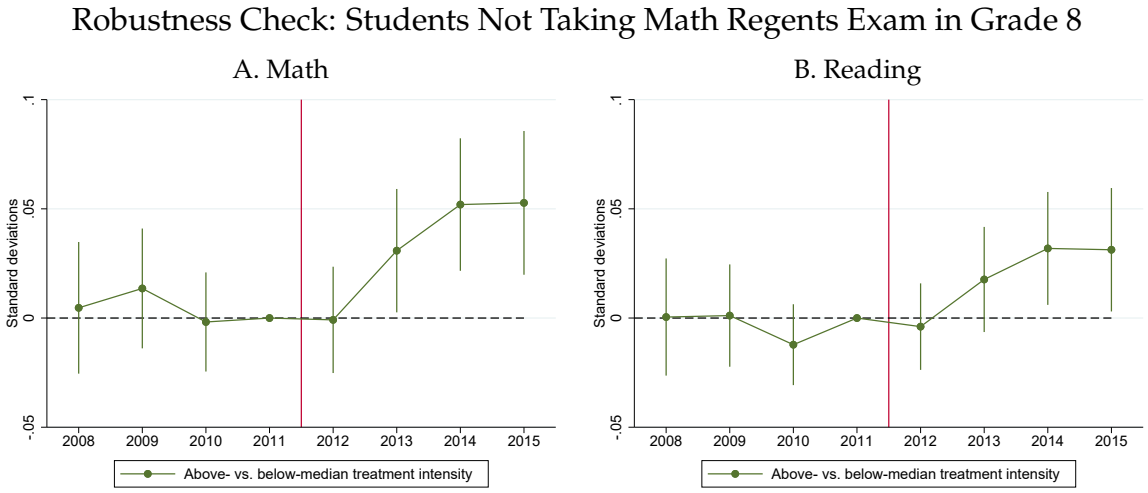

Figure E13. This figure shows the effects of the 2012 discipline reform on student math and reading achievement in a sample that excludes students who go on to take the Regents exam in math as 8th graders, or who do not reach 8th grade by 2015. Each panel plots the estimated treatment effects ρ_k from Equation 1. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in subject-grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Correlation between Current and Lagged Test Scores, Grades 6-8

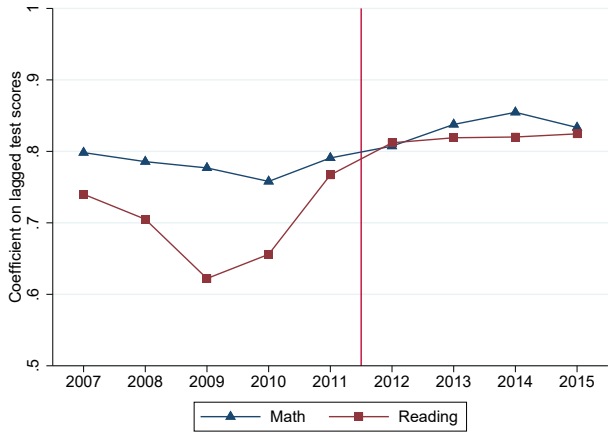

Figure F1. This figure plots the correlation between current and lagged test scores over time. For each year, we regress standardized test scores on last year's standardized test scores. We then plot the coefficients for math in blue and the coefficients for reading in red. The vertical red line indicates the timing of the switch to the Common Core testing regime. Data are from the New York City Department of Education, and include students from grades 6 through 8.

There are two reasons why the switch to the new exam is unlikely to be driving our treatment effects. First, mechanical effects from the switch would produce a sharp change in 2012, but our treatment effects gradually ramp up over the post-reform period. Second, Figure F1 shows that the correlation between last year's test score and this year's test score for a given student remains stable before and after the change in testing regimes. In other words, relative exam performance under the old testing regime (2011) predicts relative exam performance under the new testing regime (2012) just as well as past performance predicts current performance under the same regime. This is not compatible with a story in which our results are driven by tests measuring something different from 2012 onward.

A final possibility is that the change in tests could have motivated changes in teaching methodology, although the evidence from Figure F1 that the new tests measure the same thing makes this less likely. It is hard to rule this out formally, but there are two reasons to suspect that it is unlikely. First, it is not clear why such a change would be strongly correlated with pre-period suspension use, which is what we use to determine our treatment groups. Second, it is more likely that the impact would be correlated with the level of a child's test score. In this case, we would expect that rebalancing on test earlier test scores would change the results meaningfully, but Figure E4 shows that this is not the case. In fact, rebalancing on test scores has virtually no impact on our results.

Change in district leadership

In January 2014, Bill de Blasio took over as Mayor after running a campaign focused on social justice issues, including what he described as overly-harsh school discipline. He

immediately appointed Carmen Fariña as School Chancellor. Together, they advocated for further reductions in suspension rates.

This change in leadership cannot be driving our results, although it may be amplify them from 2014 onward. First, de Blasio and Fariña didn't make revisions to the discipline code until February 2015, only two months before exams for the 2014 school year (Decker and Snyder, 2015). These revisions cannot explain the test score gains we see in 2013 or the school culture improvements we see from 2012. They are also unlikely to have affected 2014 outcomes. Furthermore, our analysis in Section 6 reveals no evidence that Blasio and Fariña targeted high-suspension teachers and principals whose views on discipline were less progressive than their own. Nor is there evidence of extra resources being funneled into high-suspension schools. Nonetheless, there does remain the possibility that their public messaging could have induced schools to reduce suspension use on their own.

G Teacher Composition

We find no evidence that the gains from the reform were driven by High Treatment school-grades attracting better teachers, as measured by pay and experience.[39] Panel A of Figure G1 plots yearly treatment effects on teacher salaries, and Panel B plots the coefficients for teacher experience. In both, there is a secular decline in the High Treatment group relative to the Low Treatment group, but no change in the relationship after the reform.

H Direct Effects on Suspended Students

In Section 6.1, we argued that direct effects on suspended students would have to be very large if they were to explain the gains we see on average from this reform. To provide a concrete comparison, we now estimate an upper bound for short-term direct effects.

H.1 Empirical Design

Our methodology compares two similar groups of students: those who were suspended just before each standardized test, and those suspended just after. Our core insight here is the test scores of this latter group could not have been affected by their suspensions.

We first calculate the time in months between the start of each suspension and the 3-5 day testing window for each exam. Then we use Equation 8 to compare the scores of

[39]Teacher salary and years of experience are highly correlated in New York City. Starting salary varies with prior experience, degrees earned, and academic coursework, but then increases each year with experience. See https://www.schools.nyc.gov/careers/working-at-the-doe/benefits-and-pay.

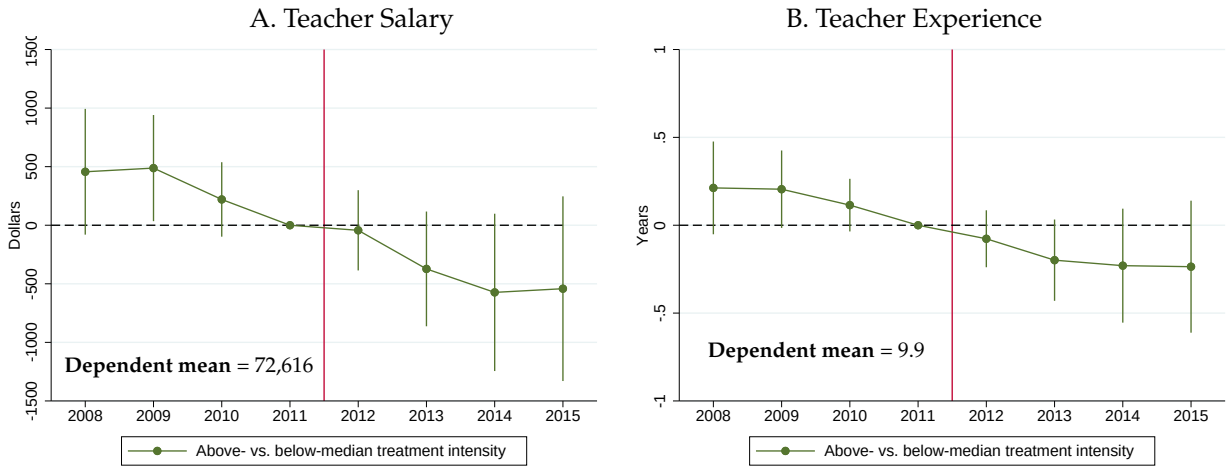

Figure G1. This figure shows the effects of the 2012 discipline reform on teacher salary and experience. The green lines plot the estimated treatment effects ρ_k from Equation 3. Each point measures the difference between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

students suspended before and after the exam. We exclude those suspended during the testing window (time 0) because we do not observe the exact exam date for each student.

$$y_{ijt} = \underbrace{\alpha_j + \gamma_t + \delta_i}_{\textbf{Fixed effects}} + \underbrace{\sum_{k \neq 0} \beta_k \mathbb{1}\left(S_{ikt} \geq 1\right)}_{\textbf{Monthly suspension indicators}} + \epsilon_{ijt} \tag{8}$$

The coefficients on the suspension indicators compare the test scores of students who are suspended in month k (relative to the exam) to students who are not. This group of students who are not suspended in month k contains both those who are not suspended at all that year, and those who are suspended, but not in that month. The comparisons are conditional on student, school-grade, and school-year fixed effects.

The difference between β_{-1} and β_1 provides our upper bound on the short-term direct impact of suspension. Mechanically, the comparison of the two coefficients captures the difference in scores between students who are suspended in the month prior to the exam, and those suspended just following the exam. Part of this difference reflects the causal effect of suspension on student achievement. However, it also captures the effect of any shock that led that student to misbehave, and any effect of the misbehavior itself. Examples of such shocks could include parental divorce or the incarceration of a family member. Nonetheless, $\beta_1 - \beta_{-1}$ is an upper bound for the short-term impact of suspension provided that such shocks *lower* students' test scores on average.

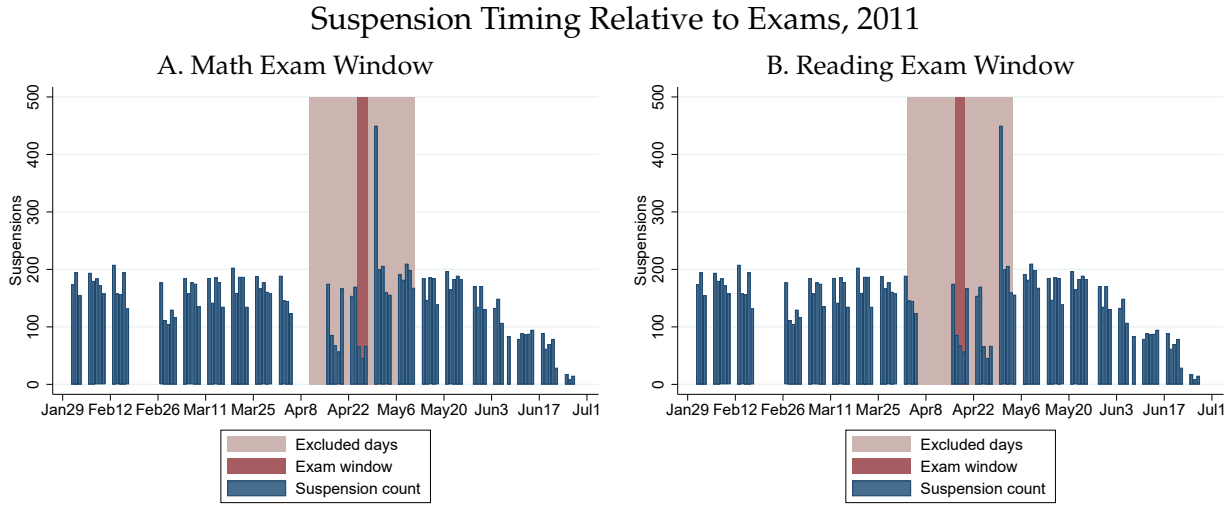

Figure H1. This figure shows the relationship between daily suspension rates and the exam windows in the 2011 school year. The blue bars show the number of suspensions issued each day to students in grades 6-8. The dark red shaded regions are the exam windows for math (Panel A) and reading (Panel B). The light red shaded regions are within two weeks of the exam window, and are excluded from the analysis. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Bunching of suspension dates

Our identification requires local variation in the timing of suspensions to be as good as random, conditional on student behavior. However, the suspension rates in Figure H1 show some evidence of manipulation: there are fewer suspensions than usual during exam windows, followed by spikes immediately after.[40] This is consistent with, for instance, schools postponing suspensions to avoid interfering with standardized testing. We therefore exclude students who are suspended between two weeks before and two weeks after the testing window. As an example, "month 1" includes suspensions occurring from two weeks after the exam to one month and two weeks after the exam. Outside of this exclusion period, we assume that decision-making about suspensions does not systematically and sharply change between the periods before and after the test.

Results of the bounding exercise

Figure H2 plots the coefficents on the monthly indicators in Equation 8. Receiving a principal's suspension has no more than a 0.03 standard deviation causal impact on a student's math score later that year, since the 95 percent confidence interval for $\beta_1 - \beta_{-1}$ ranges from -0.014 to 0.03 standard deviations. There is a much clearer discontinuity after the exam window for superintendent's suspensions, but even those longer punishments for more serious infractions have at most a 0.12 standard deviation impact. We obtain qualitatively

[40]Suspensions spike twice after the testing window for reading, with the larger spike coming on Day 9. The second spike corresponds to the end of the testing window for math a week later. See Appendix Figure I14 for daily suspension charts and exam dates for other years.

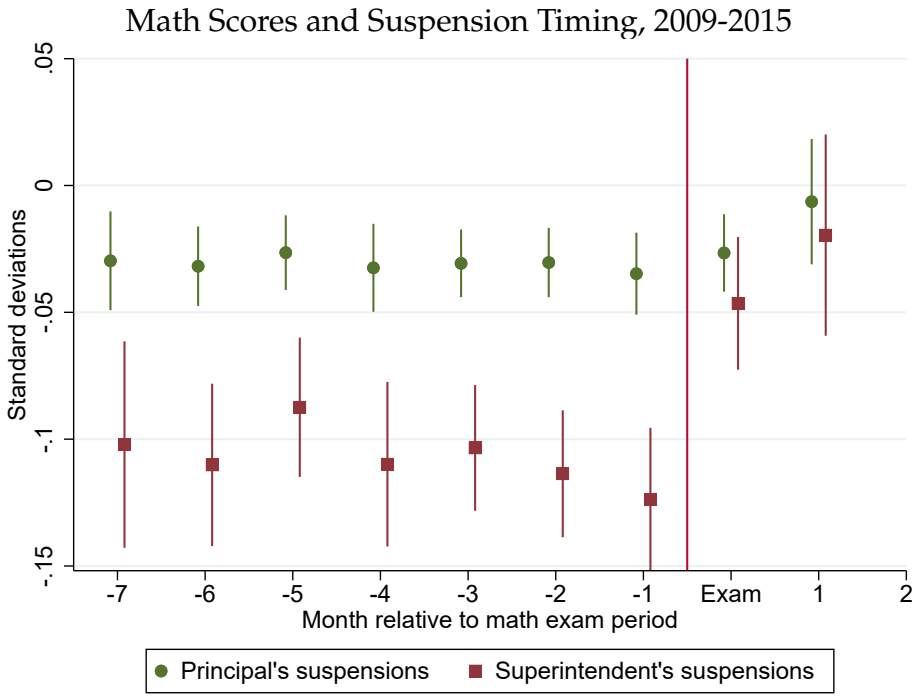

Figure H2. This table shows the monthly coefficients β_k from Equation 8. These coefficients compare the test scores of students who are suspended in month k to students who are not, conditional on year, individual student, and school-grade fixed effects. Estimates for principal's suspensions are indicated by green circles and superintendent's suspensions by dark red squares. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered by student. Data are from the New York City Department of Education, and include students from grades 6 through 8.

similar results for reading (see Appendix Figure I15).

These estimates are very small relative to the benchmarks in Section 6.1. If direct effects do not compound over time, at most 0.4 percent of the aggregate gains by 2014 can be explained by the elimination of suspensions for disorderly behavior. If the reform drove the entire reduction in suspension rates for all infractions, this rises to 2.8 percent.[41]

The main limitation of our analysis using the sharp timing of each suspension is that we cannot capture the long-term direct impact of suspensions. Our aggregate test score gains take three years to reach 0.05 standard deviations, but our estimates of direct effects are for test scores in the same year. It is possible that effects compound over time if students fall further behind as new material builds on previous material, or if psychological costs and stigma from suspension take time to affect test scores. However, direct effects would have to grow by a multiple of between 3.6 and 28 to explain even 10 percent of the aggregate test score gains.[42] We view this as unlikely.[43] Moreover, compounding

[41] We are extrapolating from the effect of the *marginal* suspension to the effect of the *average* suspension in these calculations. It is not clear ex-ante which effect should be larger.

[42] The multiple depends on whether the reform is responsible for declines in non-Level-2 suspensions.

[43] Fixed effect regressions (see Appendix A) show that students do worse than usual in years with a suspension. Appendix Figure I11 shows that they also do worse in future years, but the size of the effect di-

direct effects would create a positive gradient between treatment effects and individual suspension risk, which we do not observe in our heterogeneity analysis.

I Supplementary Tables & Figures

Table I1. Levels of Disciplinary Infractions and Suspension Length

Infractions	Description	Mean Length	Max. Length
Level 1	Uncooperative/non-compliant behavior	–	–
Level 2	Disorderly behavior	2.8 days	5 days†
Level 3	Disruptive behavior	3.4 days	10 days
Level 4	Aggressive or injurious/harmful behavior	8.3 days	1 year
Level 5	Seriously dangerous or violent behavior	27.0 days	1 year

† The discipline code reform in 2012 eliminated suspensions for Level 2 infractions.

Table notes. This table shows the five levels of infractions defined by the NYCDOE's discipline code (2008, 2011, 2012, 2013, and 2015 editions). In rare cases, principal's suspensions still occur for Level 1 or Level 2 infractions after 2012 if a student receives more than three classroom removals in the same semester. Mean suspension lengths are calculated for students in grades 6-8 between 2006 and 2011, and treat suspensions of "Between 6 Months and 1 Year" as 90 days (half of a school year). Median suspension lengths are shown in Appendix Figure I1. Appendix Table I2 lists the most common infractions that lead to suspension.

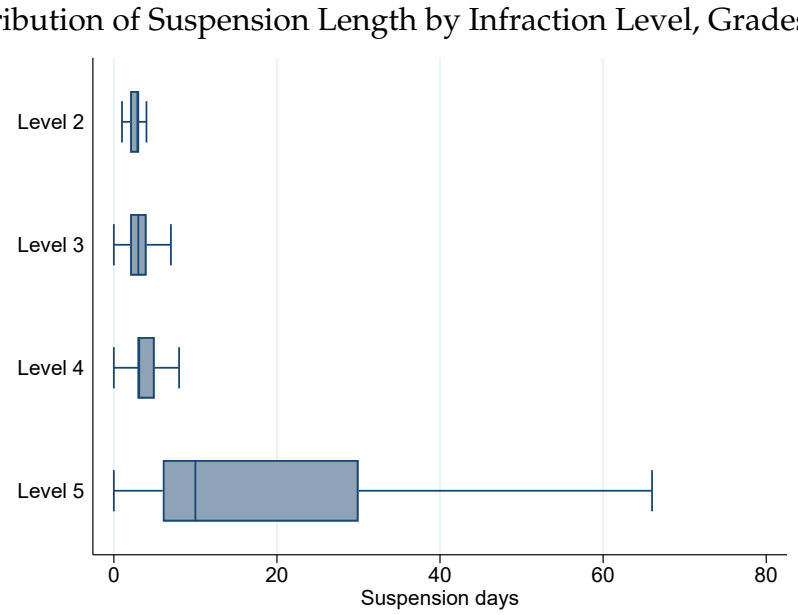

Figure I1. This figure shows the distribution of suspension length by infraction level for students in grades 6 through 8 between 2006 and 2011. Each box represents the interquartile range (IQR) and is split by a vertical line at the median. Whiskers extend from either side with length 1.5×IQR, but are truncated at zero. Data are from the New York City Department of Education.

minishes. Even though these regressions imperfectly measure the causal relationship between suspensions and test scores, we would expect larger effects at higher lags if direct effects compounded over time.

Table I2. Most Common Infractions Leading to Suspension

All Infractions		
Level	Count	Description
4	43,447	Physically aggressive behavior
3	15,402	Shoving, pushing or other similar behavior
3	14,397	Insubordination
4	10,803	Reckless behavior with risk of serious injury
4	9,865	Intimidating or bullying behavior
4	6,101	Coercing or threatening violence

Level 2 Infractions		
Level	Count	Description
2	3,847	Profane or abusive language or gestures
2	2,955	Persistent non-compliance
2	400	Lying to school personnel
2	283	Misusing property belonging to others
2	199	Smoking
2	175	Disruptive behavior on school bus

Table notes. This table shows the most common infractions that led to suspension for students in grades 6 through 8 from 2006 to 2011. The top panel shows all infractions; the lower panel limits to Level 2 infractions, which are prohibited by the 2012 reform. Data are from the New York City Department of Education.

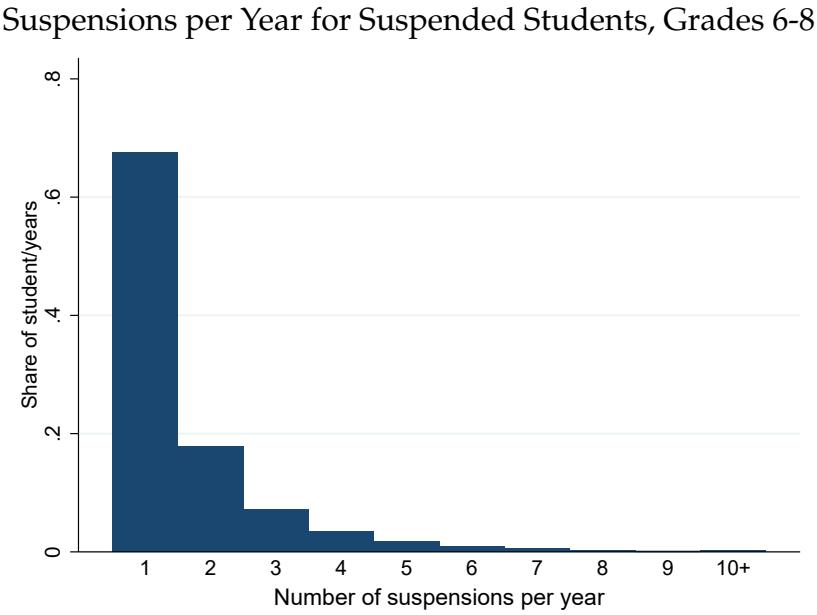

Figure I2. This figure shows the distribution of suspensions per year for all students with at least one suspension in that year, pooled across 2006-2011. Each observation is a student-year combination. Data are from the New York City Department of Education.

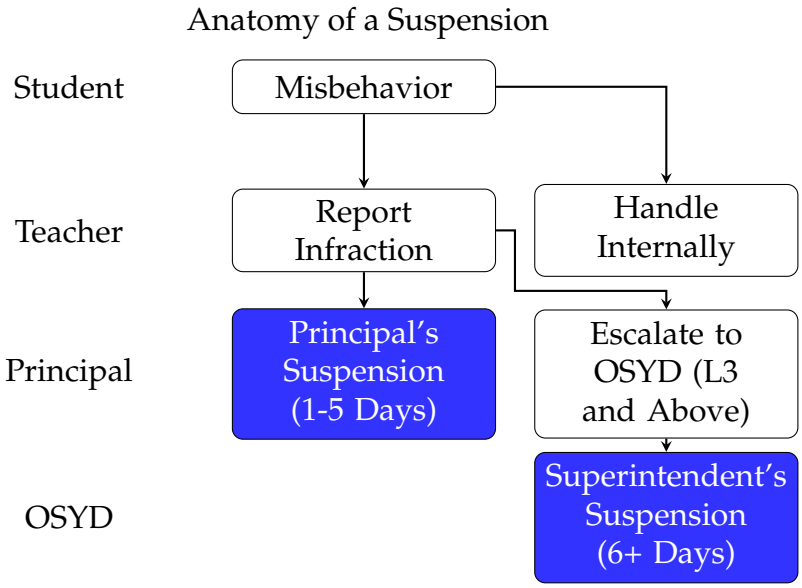

Figure I3. This figure provides a simplified outline of the decision points that lead to a suspension. Note that some disruptive behavior does not end up being a violation of the discipline code. Similarly, not every infraction results in a suspension; schools choose from a range of guidance and disciplinary interventions specified in the discipline code, and principals may decide not to formally record the infraction.

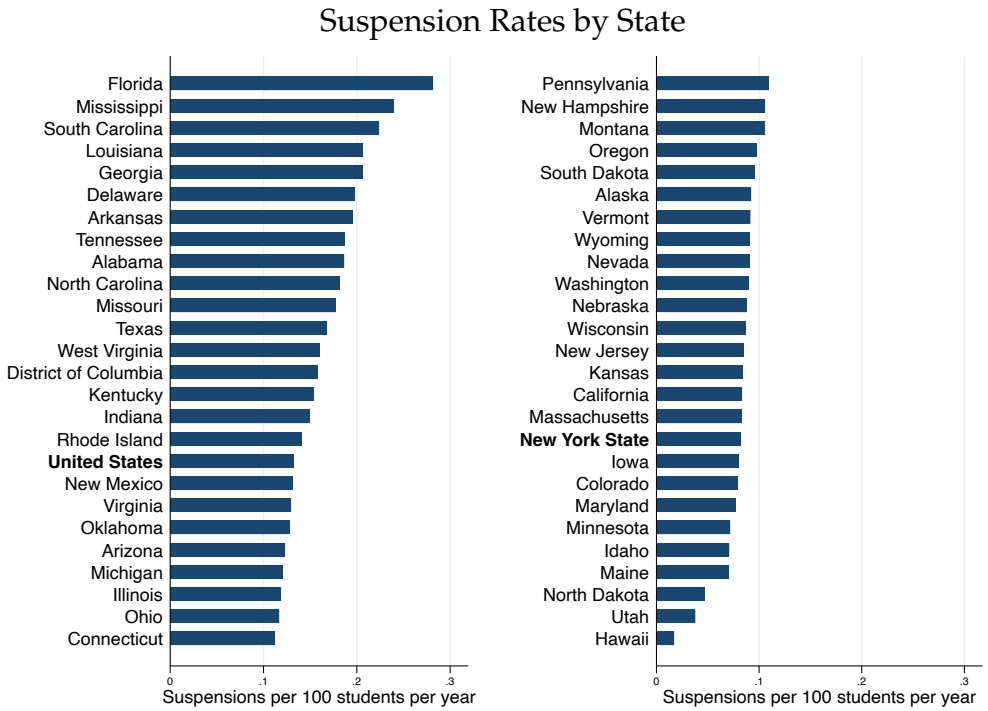

Figure I4. This figure shows suspension rates by state for all public school students in all grades in 2011-12. Data are from the United States Department of Education Civil Rights Data Collection.

Figure I5. This figure provides a regression-adjusted version of the suspension rates for the below-median treatment intensity group (Low Treatment) and our above-median treatment intensity group (High Treatment) in Figure 4. Specifically, we replace the outcome variable in Equation 1 with Level 2 suspensions instead of math scores. The results mirror the version without regression-adjustment.

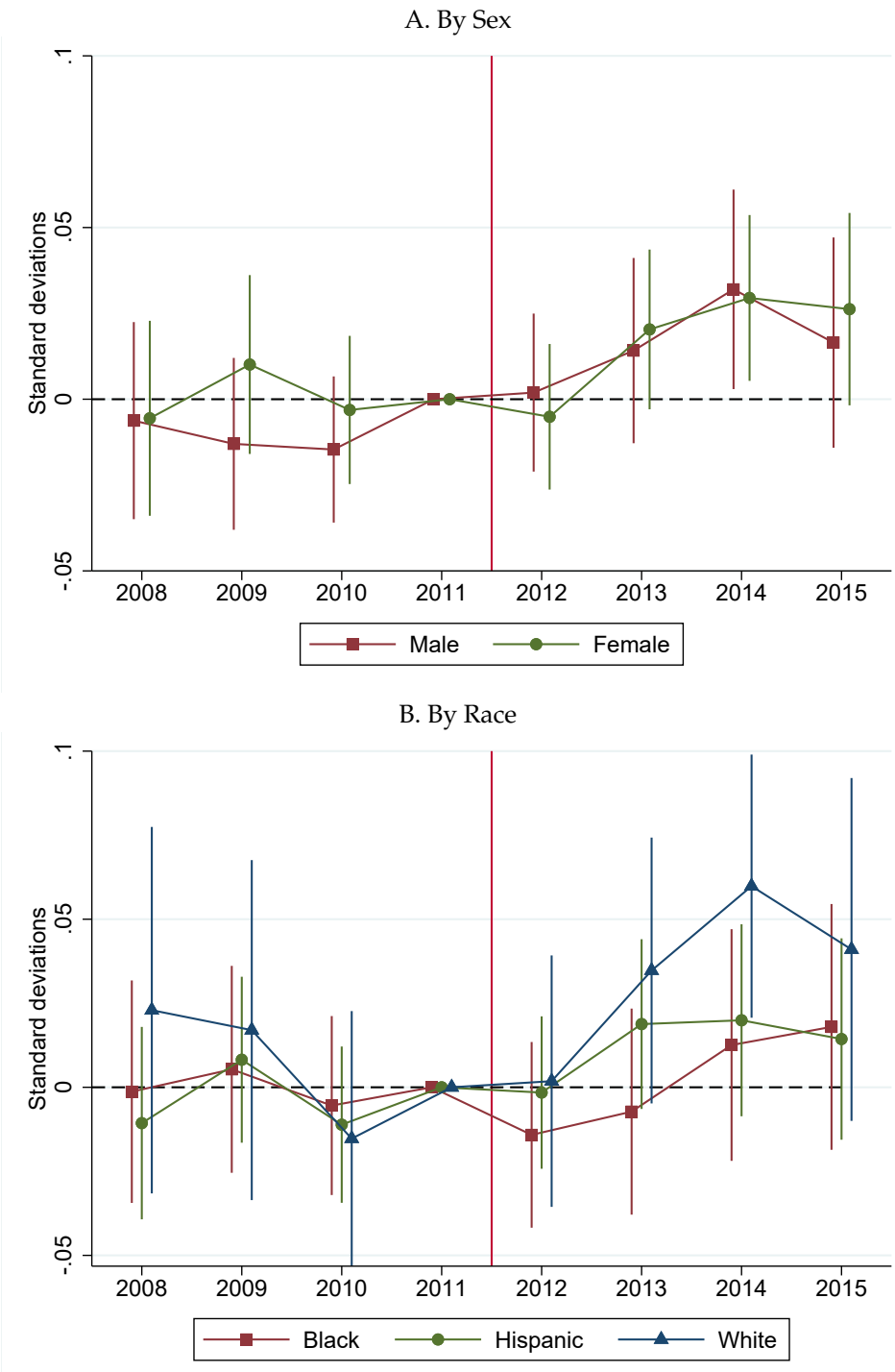

Figure I6. This figure shows the effects of the 2012 discipline reform on student reading achievement for specific demographic subgroups. Panel A plots the estimated treatment effects ρ_k from Equation 1 separately for boys and girls. Panel B plots estimated treatment effects separately for black, Hispanic, and white students. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

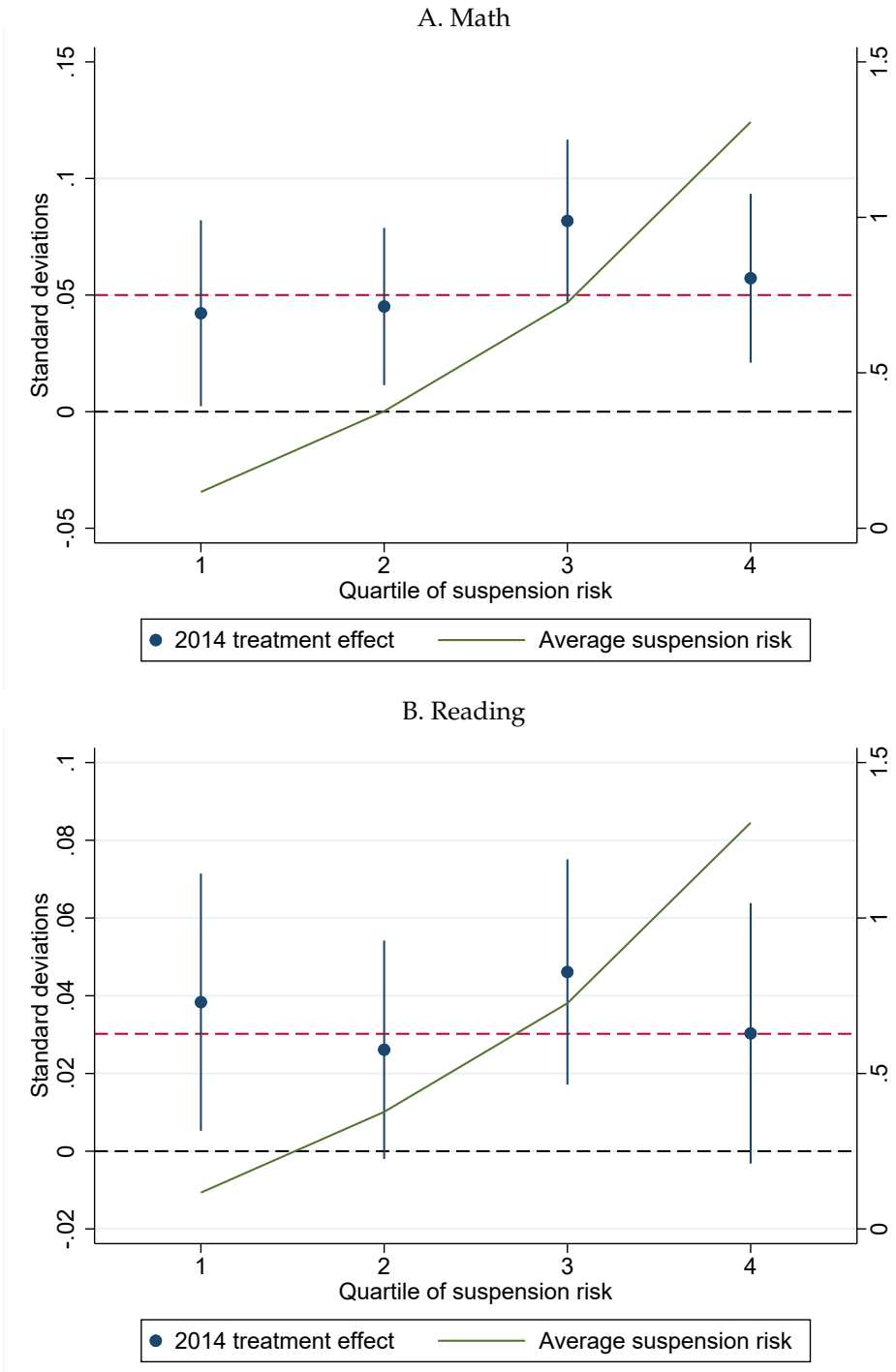

Figure I7. This figure shows the effects of the 2012 discipline reform on test scores for students in different quartiles of predicted suspension risk for Level 2 infractions only. Panel A plots the estimated treatment effects from Equation 1 in 2014 only on math scores in each risk quartile. Panel B plots estimated treatment effects in 2014 on reading scores in each risk quartile. The vertical blue lines show 95 percent confidence intervals. The red dashed line shows the estimated treatment effect in 2014 for the full sample. The green line shows actual suspension rates for Level 2 infractions within each quartile over the 2008-2011 period, measured on the right axis in suspensions per hundred students. Data are from the New York City Department of Education, and include students from grades 6 through 8.

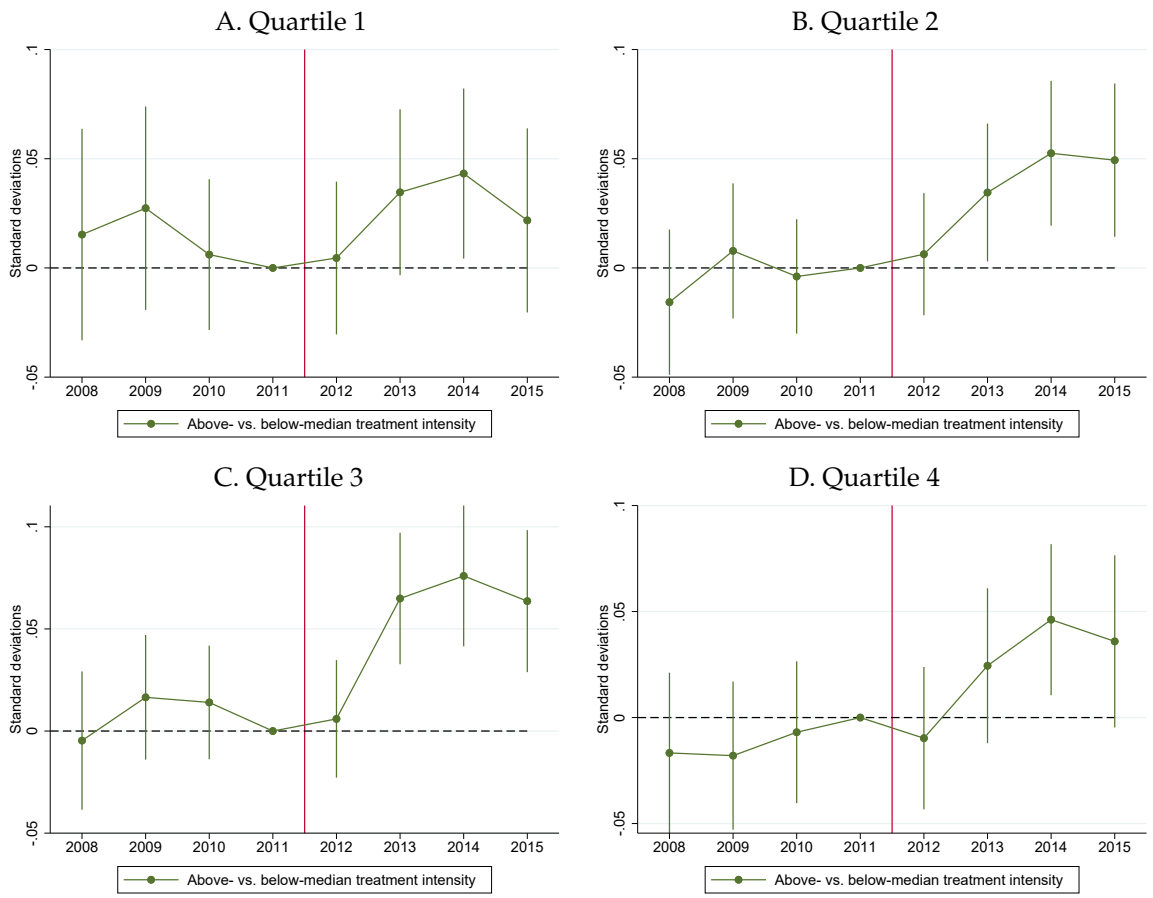

Figure I8. This figure shows the effects of the 2012 reform on student math achievement by quartile of predicted suspension risk. Each panel plots the estimated treatment effects ρ_k from Equation 1 for a given risk quartile. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

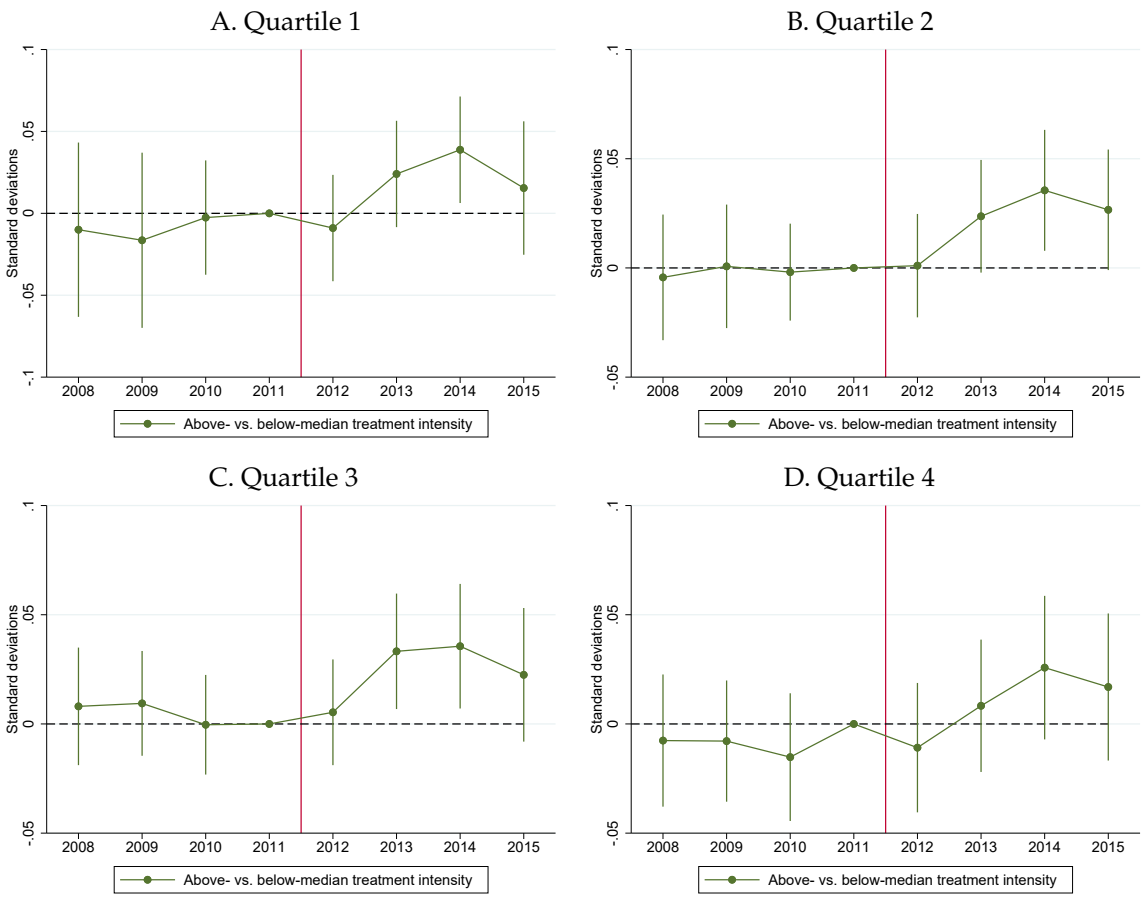

Figure I9. This figure shows the effects of the 2012 reform on student reading achievement by quartile of predicted suspension risk. Each panel plots the estimated treatment effects ρ_k from Equation 1 for a given risk quartile. Each point measures the gap in test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects and demographic controls. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Table I3. Survey Questions on School Culture

Student Survey Questions	**Availability**
Respect	2008-2013
• Most students at my school treat adults with respect.	
Behavior	2008-2015
• Students threaten or bully other students at school.	
• Students get into physical fights at my school.	
• Students use alcohol or drugs while at school.	
• There is gang activity at my school.	
Safety	2008-2015
• I am safe in my classes.	
• I am safe in the bathrooms, hallways, and locker rooms at my school.	
• I feel safe on school property outside my school building.	
Teacher Survey Questions	**Availability**
Respect	2008-2013
• At my school, most students treat adults with respect.	
Behavior	2008-2013
• Students in my school are often threatened or bullied.	
• Students' use of alcohol and illegal drugs in school is a problem at my school.	
• There are conflicts at my school based on race, color, creed, ethnicity, national origin, citizenship/immigration status, etc.	
• Gang activity is a problem at my school.	
Safety	2008-2013
• Order and discipline are maintained at my school.	
• I am safe at my school.	
• Crime and violence are a problem at my school	

Table notes. This table displays the survey questions that we use to measure respect, behavior, and safety. The right column indicates the years in which data for each group of questions is available. The exact presentation of these questions changed slightly over time. In particular: (1) the order of responses changed for some questions (strongly agree = 1 vs. strongly disagree = 1); (2) some questions had small changes to wording that didn't affect their meaning; and (3) the order of questions within the survey varied.

Changing Racial Discipline Gaps and Test Score Improvements, Grades 6-8

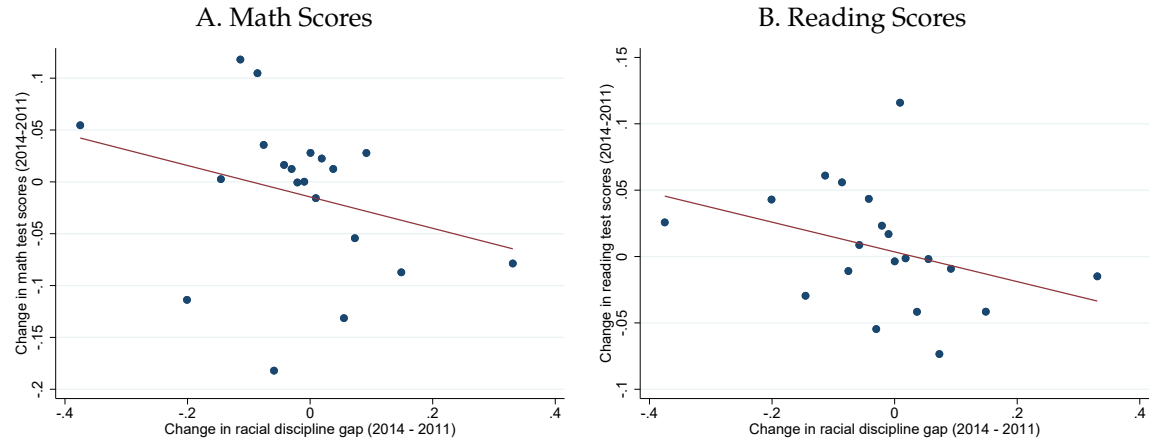

Figure I10. This figure shows the relationship between test score improvements and changes in the racial discipline gap. We define the racial discipline gap as the suspension rate for black/Hispanic students minus the rate for white students. For each school-grade, we calculate the change in test scores and the change in the racial discipline gap between 2011 and 2014. Panel A shows a binned scatterplot of these changes for math, and Panel B shows the same for reading. Test scores are standardized within the sample in subject-grade-year cells. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Suspension and Test Scores Over Time, Grades 6-8

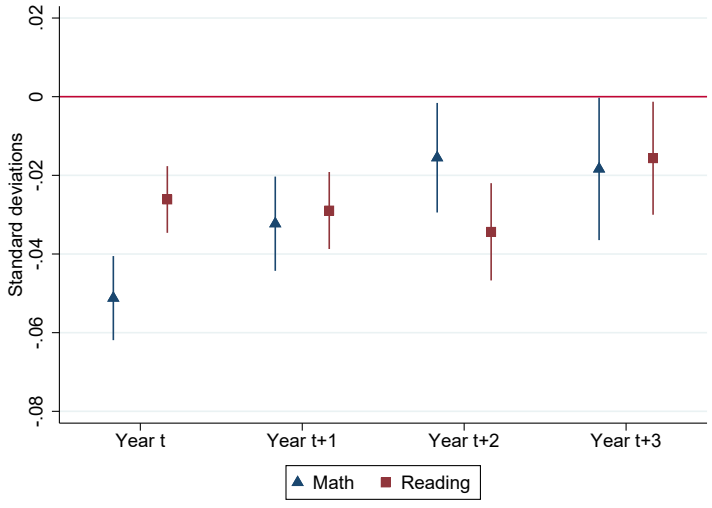

Figure I11. This figure plots the relationship between suspension in year t and test scores in years t to $t+3$, conditional on student, school-grade, and year fixed effects. Each point is the coefficient from a separate regression, estimated on data prior to the 2012 reform. The vertical bars show 95 percent confidence intervals. Test scores are standardized in subject-grade-year cells. Data are from the New York City Department of Education, and include students from grades 6 through 8.

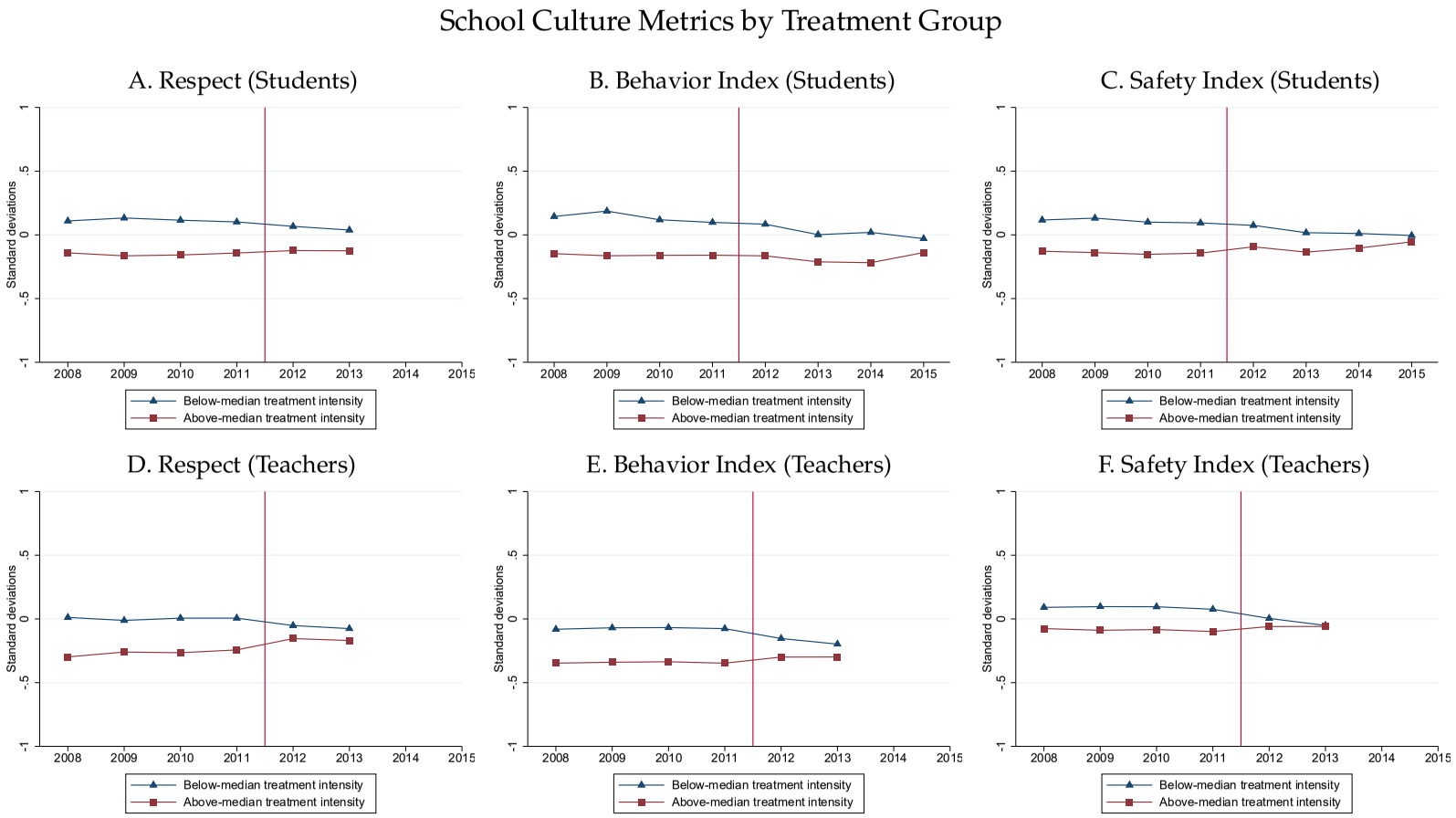

Figure I12. This figure shows average levels of our standardized culture metrics in each treatment group over time. Panels A-C plot student responses and Panels D-F plot teacher responses. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Treatment Effects on School Culture versus Reading Score Gains, Grades 6-8

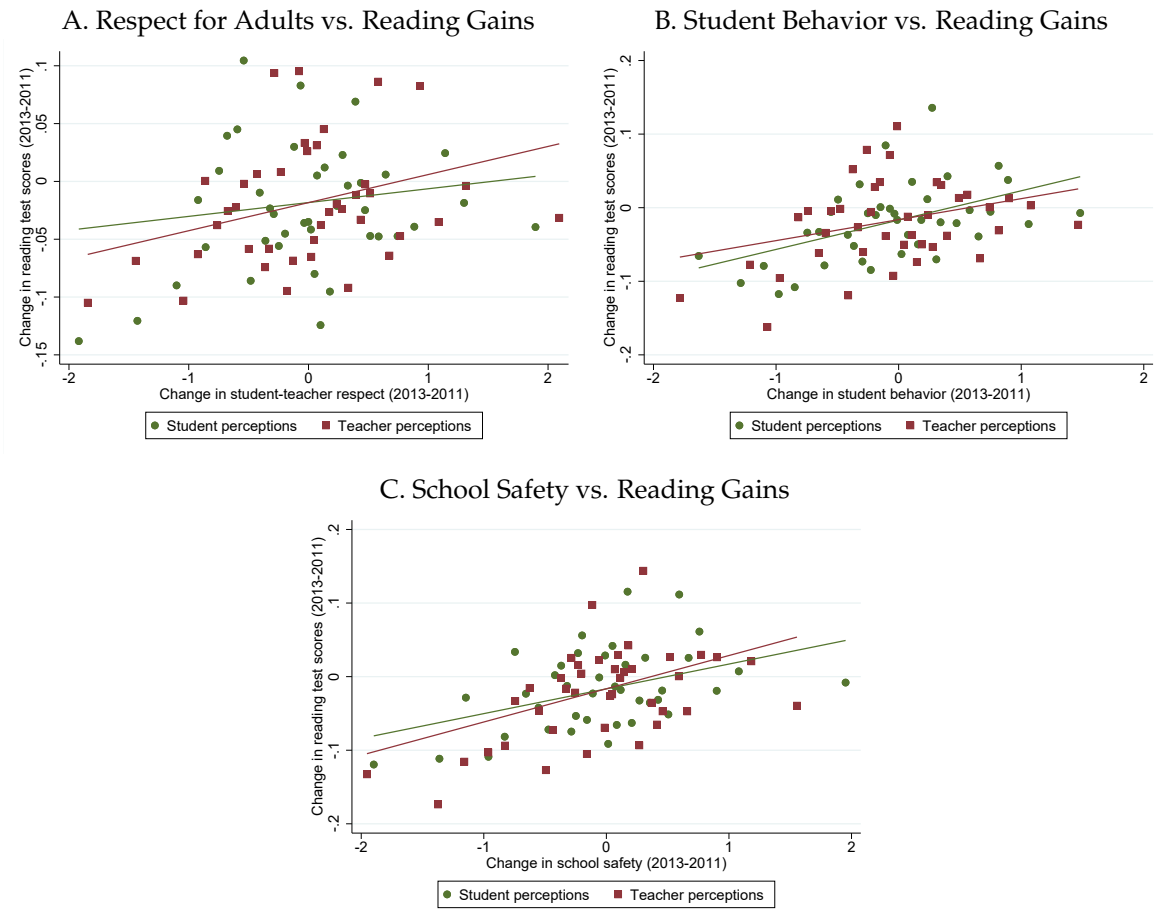

Figure I13. This figure shows the relationship between improvements in culture and test score gains in reading at the school-grade level. In Panel A, we calculate the changes in student and teacher respect between 2011 and 2013 in each school-grade and create binned scatterplots against the corresponding changes in reading test scores. Bins of student responses are in green and bins of teacher responses are in dark red. Panel B repeats the same analysis for perceptions of student behavior, and Panel C does the same for safety. Culture metrics are standardized within each year. Test scores are standardized within the sample in grade-year cells. Standard errors in all regressions are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

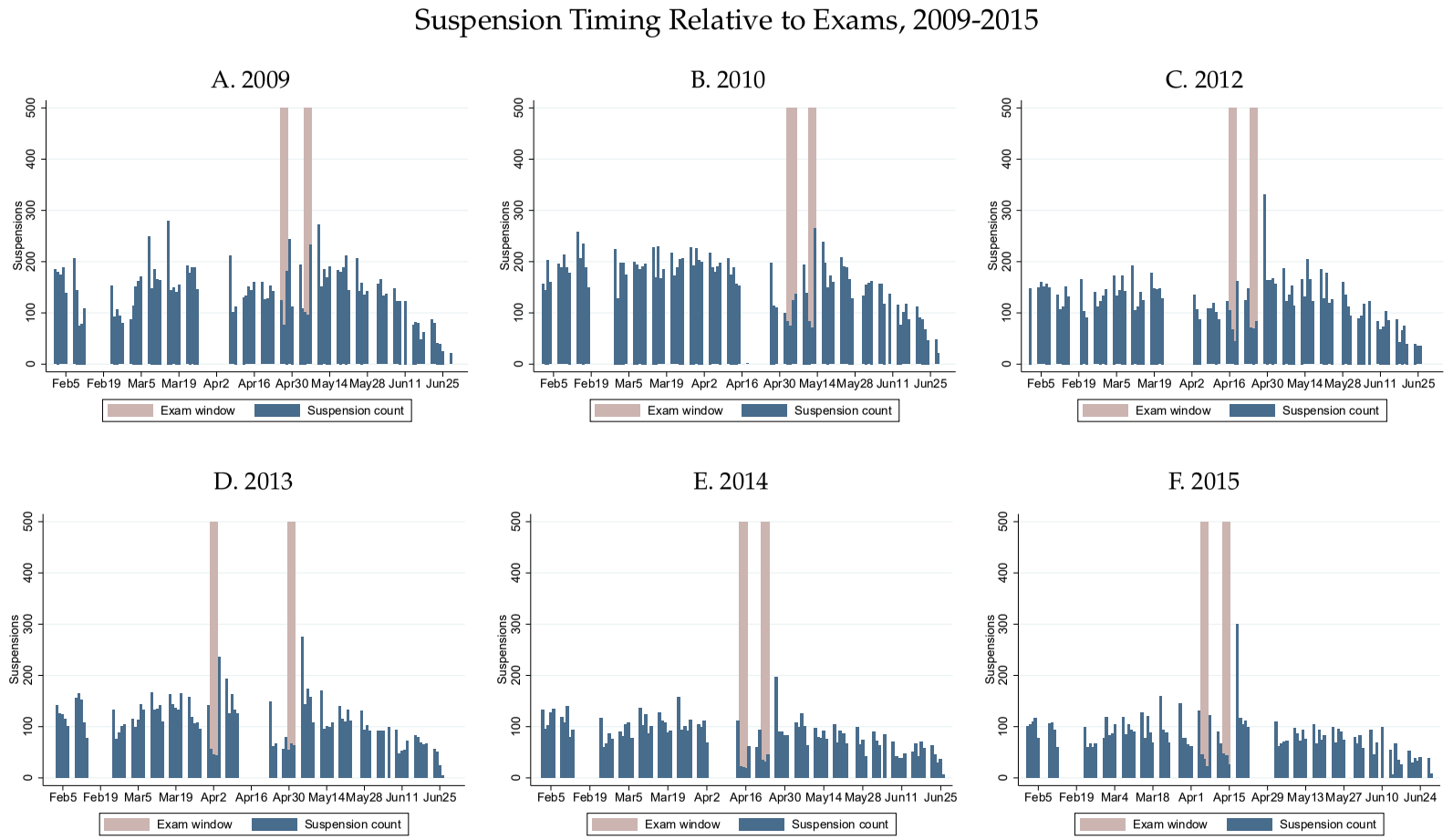

Figure I14. This figure shows the relationship between daily suspension rates and the math and reading exam windows in each school year. Suspension timing for 2011 is shown in Figure H1. The blue bars show the number of suspensions issued each day to students in grades 6-8. The light red shaded regions are the exam windows for math and reading. Data are from the New York City Department of Education, and include students from grades 6 through 8.

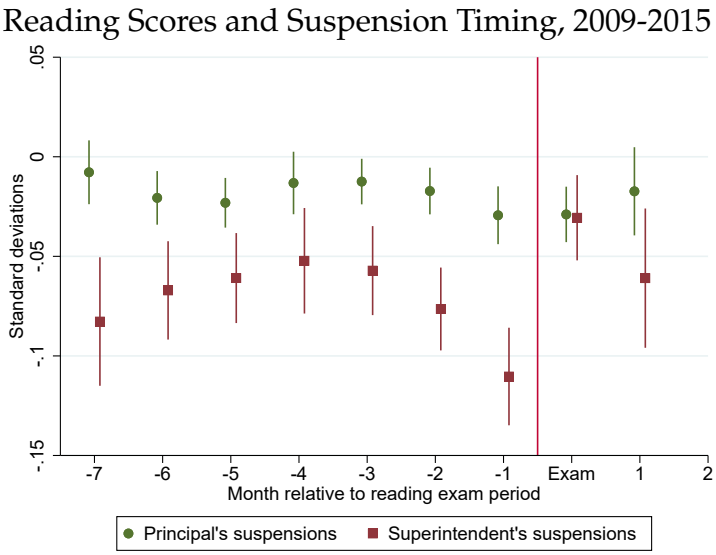

Figure I15. This table shows the monthly coefficients β_k from Equation 8. These coefficients compare the test scores of students who are suspended in month k to students who are not, conditional on year, individual student, and school-grade fixed effects. Estimates for principal's suspensions are in green and superintendent's suspensions are in dark red. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered by student. Data are from the New York City Department of Education, and include students from grades 6 through 8.

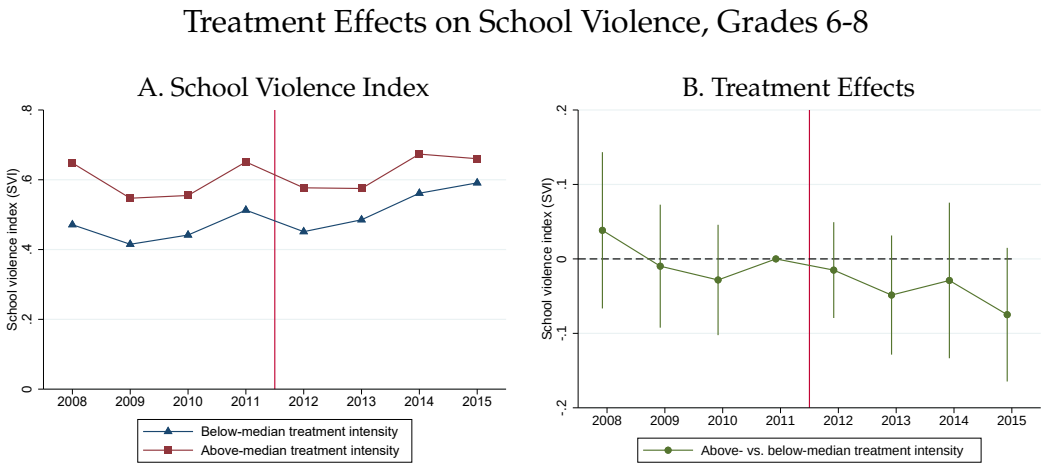

Figure I16. This figure shows the effects of the 2012 discipline reform on an index of school violence recorded through VADIR. Panel A shows average levels of the School Violence Index (SVI) and Panel B plots estimated treatment effects from Equation 3. Each point measures the gap in the SVI between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

Discipline Reform and Math and Reading Achievement, No Controls

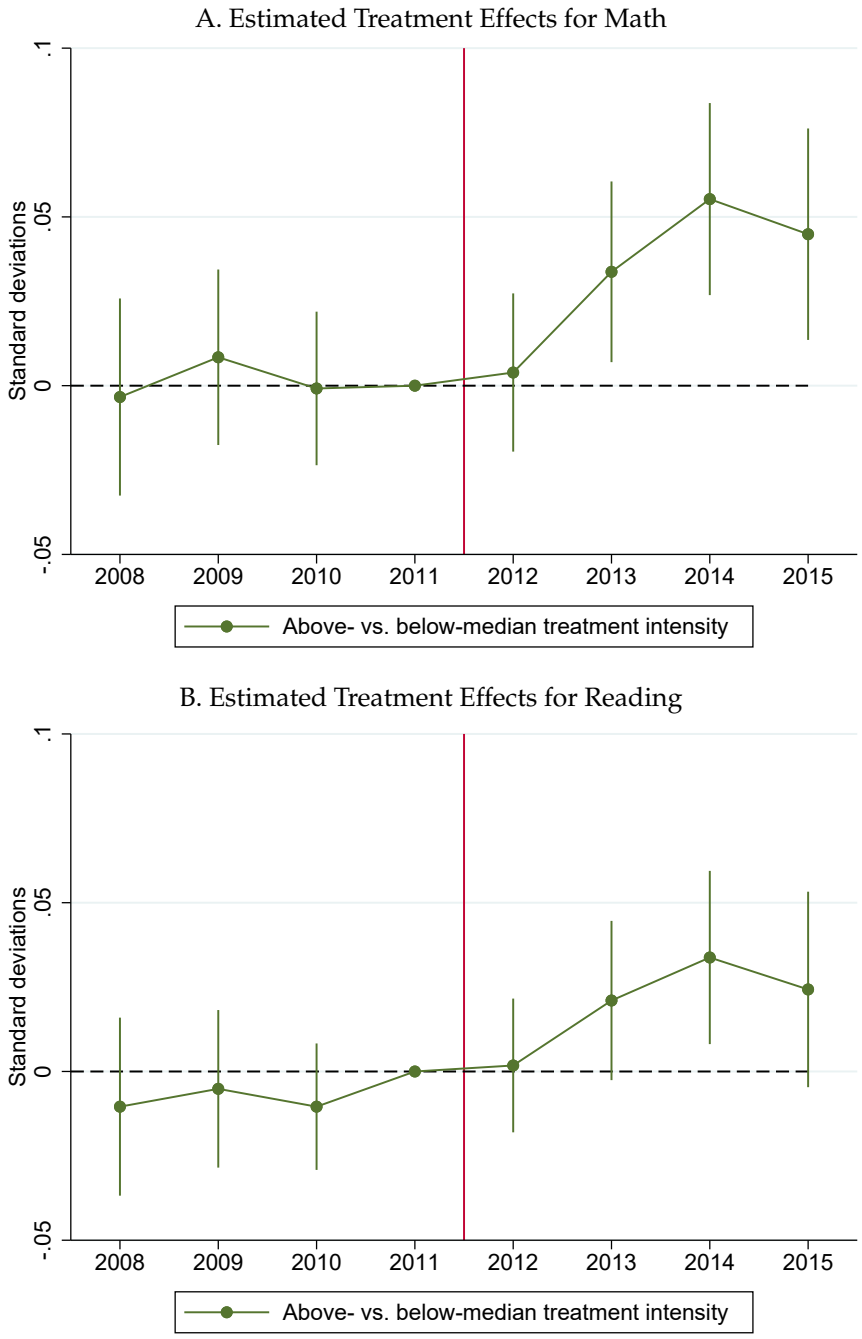

Figure I17. This figure shows the effects of the 2012 reform on math and reading achievement. Panel A plots the estimated treatment effects for math, ρ_k, from Equation 1. Each point measures the gap in math test scores between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. However, time-varying controls are omitted here. Panel B does the same for reading. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Test scores are standardized within the sample in grade-year cells. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.

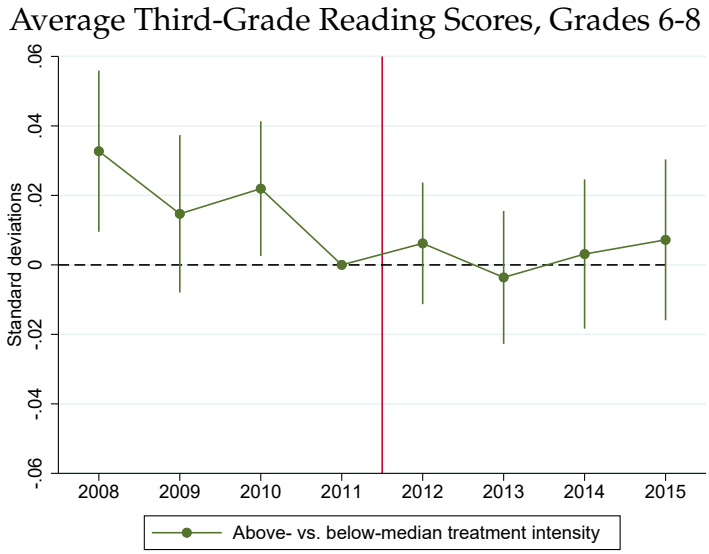

Figure I18. This figure shows the effects of the 2012 reform on school-grade ability, as measured by average reading test scores from grade 3. The green lines plot the estimated treatment effects ρ_k from Equation 3. Each point measures the difference between the High and Low Treatment groups relative to 2011, conditional on year and school-grade fixed effects. The vertical bars show 95 percent confidence intervals, and the red line indicates the timing of the reform. Standard errors are clustered at the school-grade level. Data are from the New York City Department of Education, and include students from grades 6 through 8.