NBER WORKING PAPER SERIES

FROM RETRIBUTIVE TO RESTORATIVE: AN ALTERNATIVE APPROACH TO JUSTICE

Anjali Adukia
Benjamin Feigenberg
Fatemeh Momeni

Working Paper 31675
http://www.nber.org/papers/w31675

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023

For helpful feedback, we thank Roseanna Ander, Hellen Antonopoulos, Monica Bhatt, Sandy Black, Jessika Bottiani, Jon Guryan, Ariel Kalil, Jens Ludwig, Ben McKay, Martha Minow, Dick Murnane, Steven Raphael, Stephen Raudenbush, Jean Sack, Anita Wadhwa, Chezare Warren, staff members at the Chicago Public Schools, and seminar participants at APPAM, the NBER education, NBER children's, and NBER Summer Institute crime meetings, the National Academy of Education, the Spencer Foundation, the William T. Grant Foundation, University of Illinois at Chicago, and The University of Chicago. For excellent research support, we thank David Arbelaez, Bryant Cong, Emily He, Juan Miguel Jimenez, Jiayu Kang, Farah Mallah, Eleni Packis, Puja Patel, Priyal Patil, Sarah Permut, and Jalisia Taylor-Singleton. For financial support, we thank the Becker Friedman Institute for Economics at the University of Chicago, the National Academy of Education, Spencer Foundation, the Successful Pathways from School to Work initiative of the University of Chicago funded by the Hymen Milgrom Supporting Organization, and William T. Grant Foundation. For data acquisition, we thank Chicago Public Schools and Chicago Police Department. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

From Retributive to Restorative: An Alternative Approach to Justice
Anjali Adukia, Benjamin Feigenberg, and Fatemeh Momeni
NBER Working Paper No. 31675
September 2023
JEL No. I0,I20,I21,I24,J0,J01,J08,J18,K39

ABSTRACT

School districts historically approached conflict-resolution from a zero-sum perspective: suspend students seen as disruptive and potentially harm them, or avoid suspensions and harm their classmates. Restorative practices (RP) -- focused on reparation and shared ownership of disciplinary justice -- are designed to avoid this trade-off by addressing undesirable behavior without imparting harm. This study examines Chicago Public Schools' adoption of RP. We identify decreased suspensions, improved school climate, and find no evidence of increased classroom disruption. We estimate a 19% decrease in arrests, including for violent offenses, with reduced arrests outside of school, providing evidence that RP substantively changed behavior.

Anjali Adukia
Harris School of Public Policy
The University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
adukia@uchicago.edu

Benjamin Feigenberg
Department of Economics
University of Illinois at Chicago
University Hall room 706
601 South Morgan Street
Chicago, IL 60607
bfeigenb@uic.edu

Fatemeh Momeni
Crime and Education Labs
University of Chicago
33 N LaSalle St.
Chicago, IL 60602
fmomeni@uchicago.edu

A data appendix is available at http://www.nber.org/data-appendix/w31675

I Introduction

Classroom management and discipline represent one the hardest parts of school officials' jobs (Evertson and Weinstein, 2006; Kauffman et al., 2011). Over the last five decades, educational authorities have increasingly turned to using exclusionary discipline, with the rate of school suspensions more than doubling for Black and Latine children since 1974 (Losen, Martinez et al., 2020).[1] In school year (SY) 2012, approximately 3.5 million public school students were suspended from school, losing nearly 18 million days of instruction due to "zero-tolerance" policies (Losen et al., 2015). Being in a stricter school can lead to long-term negative consequences such as decreased educational attainment, increased misconduct, and increased likelihood of engaging with the criminal legal system (Fabelo et al., 2011; Shollenberger, 2015; Wolf and Kupchik, 2017; Bacher-Hicks, Billings and Deming, 2019).

One justification for policies encouraging more punitive discipline is the desire to prevent negative spillover effects from disruptive students that make it hard for other students to learn. This concern has led districts to approach conflict resolution from a zero-sum perspective: suspend students whom they view as disruptive in an effort to hold them accountable and potentially harm those students, or avoid suspensions and harm their classmates whose learning now may be disrupted. While educators are increasingly aware of the potential harms of suspensions, they seek concrete responses to undesirable behavior, particularly in a context where 80 percent of schools report having incidents of violence, theft, or other crimes (Griffith and Tyner, 2019; Wang et al., 2020). Indeed, over two-thirds of parents and teachers have historically offered strong support for the establishment of zero-tolerance policies to promote accountability (Public Agenda Foundation, 2004). In recent years, a small but growing movement within education has sought to find a solution that avoids this tradeoff by deterring undesirable behavior without imparting harm.

In our study, we investigate one such alternative: restorative justice (RJ) practices, which emphasize community building and restitution or restoration, as an alternative to the traditional punitive approach (Losen, Hewitt and Toldson, 2014). RJ as a philosophy emphasizes the reparation of harm between victims and offenders, engaging various stakeholders in the community through open dialogue and shared ownership of disciplinary justice with the goal of restoring (or transforming) relationships and fostering long-term reparative approaches to conflict resolution (McCold and Wachtel, 1998; Fulkerson, 2001; Karp and Breslin, 2001; McGarrell, 2001; Hopkins, 2003; González, 2012; Angel et al., 2014; Wadhwa, 2015; Winn, 2016; Augustine et al., 2018; Gregory et al., 2018; Acosta et al., 2019; Shem-Tov, Raphael and Skog, 2021; Minow, 2022). While RJ has entered U.S. public education

[1]Hereafter, we refer to Black or African American children as Black children and Latine/a/o or Hispanic children as Latine children.

systems only recently in the form of restorative practices (RP), it has quickly increased in use despite the lack of quantitative evidence on the associated costs and benefits. Studies on the impacts of RP on educational and behavioral outcomes, within or outside of schools, are informative but limited with most being correlational or descriptive.

This study provides, to the best of our knowledge, the first credible estimates of the effects of large-scale implementation of restorative practices in an educational setting. We leverage the rollout of RP programs across 73 high schools within the Chicago Public Schools (CPS) system beginning in school year (SY) 2013-2014.[2] Collectively, the 239 high schools in our study sample (including those that did not implement RP) serve over 100,000 students annually. To expand access to RP programming in schools, CPS provided training to school staff that emphasized less punitive and more reparative strategies when engaging with students (for example, developing restorative mindsets and language in school staff, creating and implementing restorative protocols and processes in response to disciplinary incidents, and strengthening student-teacher relationships). Using a difference-in-differences-style research design (based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020)), we examine how student educational and behavioral outcomes, school climate perceptions, and criminal legal system engagement respond to RP exposure.[3] Given evidence of the disparate impact of traditional disciplinary approaches on male students, Black and Latine students, and students with disabilities, we explicitly explore heterogeneity by student gender, race, engagement with special education (IEP) or a 504 plan which indicates a physical and/or cognitive disability, and English Learner status.

We find that restorative practices decreased out-of-school suspensions by 18% for high school students. We do not find evidence of corresponding increases in in-school suspensions, suggesting that students are receiving more in-school instruction time in response to policy adoption. There are two potential explanations for these findings. First, the effects may be mechanical because school administrators and teachers were instructed to reduce the frequency of suspensions. Alternatively, it may be that RP is having a positive, productive impact on teacher behavior and/or student behavior. Teachers may be changing how they interact with students, better responding to students' individual needs, and avoiding escalation. RP may teach students how to resolve conflicts more effectively, to understand their roles in conflicts, and to feel more understood by adults and their peers.

To distinguish between these alternative explanations for the measured declines in suspensions, we use person-level arrest data from the Chicago Police Department. We identify

[2]For brevity, we will refer to school years by the year in which the spring term occurs (*e.g.*, school year 2013-2014 is 2014 or SY14), following CPS convention.

[3]In additional analyses, we examine outcomes at the elementary-school level.

a 19% overall decline in child arrests, with significant decreases both during school hours and on school grounds (35%) and outside of school (15%). The documented decline in arrests is significant for both violent (18%) and non-violent (20%) offenses. This evidence suggests that the introduction of restorative practices generated meaningful changes in underlying student conduct and demonstrates that school practices may meaningfully shape socializing behaviors. This is consistent with literature indicating that schools act as socializing institutions where their disciplinary policies may reach beyond the creation of conditions for learning in the short term, sending signals to children about optimal ways to behave (Parsons, 1959; Dreeben, 1967; Bowles and Gintis, 1976).

Additionally, in accordance with the theory that RP may shift school culture, we find an improvement in student-reported perceptions of school climate, which include survey measures related to classroom behavior of peers, psychological sense of school membership, student-teacher trust, and school safety. These changes may also contribute to improvements in perceived behavior and may help to explain our finding that students enrolled in RP-adopting schools are less likely to exit the CPS system.[4]

Exploring the importance of RP program design, we show our results are driven by CPS' intensive coaching program (RP Coaching) that involved ongoing professional development of school staff throughout the school year. This helps to reconcile our conclusions with prior work indicating that less intensive training programs often do not result in substantive effects (LaLonde, 1986).

A common concern is that reduced punitiveness in the absence of behavioral change may lead to increased classroom disruption. Pope and Zuo (2020) highlight the deficiencies of simply restricting teachers from using exclusionary discipline without providing alternative tools to address misconduct. While they find that suspension reduction policies introduced in the early 2000s in Los Angeles Unified School District led to dramatic declines in suspension rates, these reforms also resulted in significant declines in academic performance, as well as increases in absences and in teacher turnover. By contrast, we do not identify significant GPA or test score changes in response to the introduction of RP. Our findings suggest that the shift toward restorative practices does not seem to have been detrimental to the learning outcomes of the broader student body, on average. Evidence of improvements in students' perceptions of classroom behavior also points against increases in classroom disruption. To more rigorously assess how (if at all) classroom disruption effects contribute to our conclusions, we employ a random forest algorithm to classify students based on their classmates' predicted suspension rates under the status quo disciplinary system. We show that differ-

[4]We speak of "perceived" behavior because it may be that students are actually behaving in an undesirable way or it may be that adults are *perceiving* them to be behaving in an undesirable way.

ences in predicted suspension rates in turn predict differences in suspension rate declines in response to the introduction of restorative practices. We isolate potential disruption effects by focusing on students who are themselves at low risk of suspension and therefore less likely to experience any suspension-related change in instructional time, and we find that any negative test score impacts are concentrated in schools with below-median predicted suspension rates. These results strengthen our determination that disruption effects appear limited in the study setting.

Finally, we investigate treatment effect heterogeneity with a focus on student race and gender, two of the strongest observable predictors of baseline exposure to suspensions and arrests. We find that Black students benefit most consistently from the introduction of restorative practices. Black males in particular, who are suspended for four times as many days as their white male peers and arrested six times more frequently at baseline, experience the largest declines in out-of-school suspension days and arrests. Correspondingly, Black male students see significant improvements in attendance (above and beyond the increase associated with reduced suspension days) and in math test scores. Treatment effect estimates indicate that the introduction of restorative practices closes the math test score gap between Black male students and their peers (non-Black male students) by 15%. By contrast, we find negative and marginally significant test score impacts for Latino male students.[5]

It is possible that Black students benefit differentially from RP because the schools they attend implement RP more effectively. To assess this possibility, we investigate treatment effect heterogeneity based on the baseline enrollment share of Black students at the school level. We find that the suspension rate declines experienced by Black students do not vary systematically with school racial composition. At the same time, point estimates suggest that math test score gains for Black students are largest in those schools where Black students are concentrated and any test score declines for Latine students are concentrated in schools with low Black enrollment shares. Definitively identifying the mechanism that underpins these patterns is beyond the scope of the present study, but our findings may be explained by the greater capacity for academic gains in those schools that are most reliant at baseline on punitive discipline and face the greatest deficits in terms of school climate perceptions and academic performance.

Taken together, our results provide evidence that an alternative approach exists that helps solve the perceived tradeoff between educators feeling the need to suspend students or else face learning disruptions within schools. Indeed, our findings suggest that no such tradeoff exists. Instead, RP has the potential to improve student perceptions of school cli-

[5]Estimates for Latino male students decrease in magnitude and are no longer significant at conventional levels once we account for differential test score missingness in response to RP adoption.

mate and reduce behavioral incidents inside and outside of school without harming academic performance, potentially improving the daily experiences of all students, regardless of their ex-ante exposure to exclusionary disciplinary practices.

The rest of the paper proceeds as follows. In Section II, we describe a conceptual framework related to how restorative practices may influence outcomes in schools. In Section III, we describe the policy setting. In Section IV, we discuss the data we use to estimate impacts. In Section V, we explain our research design. In Section VI, we discuss our findings. In Section VII, we discuss possible disruption effects as a mechanism. In Section VIII, we present treatment effect heterogeneity by student characteristics. In Section IX, we conclude.

II Conceptual Framework: Shaping Student Behavior in Schools

School officials view classroom management and discipline as an important but difficult aspect of their roles. Their goal is to create an environment that is conducive for learning (Evertson and Weinstein, 2006; Kauffman et al., 2011). This involves responding to what they perceive as being undesirable behavior. Historically, this response has taken the form of exclusionary discipline, such as suspensions, without alternative options. The advent of restorative practices offers education authorities the opportunity to avoid the tradeoff between suspending students or facing increased disruptions to learning within their schools.

Consider a simple example involving two periods where there is an incident with three main student actors. In period 1, a student exhibits undesirable behavior ("the one who harmed," or the "offender") towards another individual ("the one who was harmed," or "victim"), where there are other students who passively or actively observe the incident ("bystanders").

In period 2, the school officials respond to the undesirable behavior. There are different goals for each student actor. First, the goals for the offender are that they are held accountable and that they learn appropriate behavior for the future. Second, for the victim, the goals of any response are that they feel safe, "whole" again, and that justice has been served. Third, the goals for the bystanders are to feel safe, to learn appropriate behavior, and to be deterred from exhibiting the undesirable behavior in the future.

A common response by school officials involves exclusionary disciplinary practices, typically in the form of suspensions. However, this approach does not meet the standards for an optimal response based on the goals outlined for period 2. At best, a suspension removes the offender from a situation but neglects to impart desired behavior. It may temporarily increase the victim's feeling of safety and provide a reprieve from interacting with the offender, but it remains unclear whether they feel justice, for justice itself necessitates a sense of accountability. Victims often report that offenders need to understand the harm that they

caused in order for that offender to truly feel accountable for their actions. In the case of suspensions, if the offender is simply removed from a situation without understanding the harm they caused or how they made the victim feel, it is more difficult for them to take proper responsibility for their actions. The best-case scenario for the bystanders in the case of a punitive response in the form of a suspension is that they feel safer, there is a deterrence effect, and they possibly have a reprieve if the offender was causing disruptions to the learning or social environment.

At worst, the exclusionary response could be counterproductive to the long-term goals of school officials and perpetuate long-term harm through negative impacts on educational attainment or criminal legal system involvement (Fabelo et al., 2011; Shollenberger, 2015; Wolf and Kupchik, 2017; Bacher-Hicks, Billings and Deming, 2019).

School officials are increasingly aware of the negative consequences associated with a stricter school environment. Teachers, however, report needing concrete tools to meaningfully achieve justice and accountability in response to disciplinary situations without generating the potential harms related to exclusion.

This has led to the introduction of "restorative justice" (RJ). RJ is an approach that involves repairing harms between victims and offenders and restoring relationships, or transforming them in cases where there was not a pre-existing relationship. In RJ, the different stakeholders are engaged through open dialogue with the goal of increased perspective taking and shared ownership of disciplinary justice. The concept originated in the criminal legal system, and increasingly, school districts across the U.S. have been adopting the RJ approach to purposively shift away from the punitiveness of past policies.

RJ is typically referred to as restorative practices (RP) in the school context because it can constitute a range of practices, including restorative conversations, peer juries, and peace circles. Broadly speaking, RP should engender a sense of justice for each party involved. It can involve a conference between the offender and the victim, bringing two victims together who went through similar experiences, promoting interactions between an offender and a victim's family or friend circle, or bringing together two people who committed similar offenses. Each agent has to agree to whatever the process is; a victim will not be forced to participate if they feel that the process will re-traumatize them or if they do not want to discuss their experiences. By design, the precise structure of RJ is intentionally flexible and will vary based on the setting and situation (McCold and Wachtel, 1998; Fulkerson, 2001; Karp and Breslin, 2001; McGarrell, 2001; Hopkins, 2003; González, 2012; Angel et al., 2014; Wadhwa, 2015; Augustine et al., 2018; Gregory et al., 2018; Acosta et al., 2019; Shem-Tov, Raphael and Skog, 2021; Minow, 2022).

Concretely, consider an instance of property damage in which a student writes on a

school wall. This incident may be addressed, for example, through a peace circle, which typically entails a planned, structured meeting between the individual(s) who caused harm, the individual(s) who were harmed, and any relevant people in their communities (families and friends). In this example, the circle would typically take place between the student and the custodian, and the student would hear from the custodian about the impact of their wrongdoing. The goal would then be to repair the harm done by determining logical consequences that are fair, sensible, and directly tied to the the problematic behavior. For this example, such consequences could include the student performing community or school clean-up projects or shadowing/helping the custodian for the day. This emphasis on identifying logical consequences that can serve to promote learning and self-reflection, as opposed to employing one-size-fits-all punitive punishments, is a unifying theme of RP regardless of the precise behaviors being addressed.

In theory, a restorative approach to shaping student behavior thus provides schools with an option that allows them to hold students accountable for their actions without using exclusionary discipline and without disrupting learning within school, thus giving educational authorities the tools needed to avoid the traditional "zero-sum" approach.

III Policy Setting: Chicago Public Schools

We study the impacts of restorative practices in the context of the Chicago Public Schools (CPS), one of the largest school district in the U.S., which serves over 340,000 students annually across more than 600 schools. The population of CPS is racially and economically diverse. Of the students attending CPS in SY21, 36% identified as Black, 47% as Latine, and 11% as White, and over 63% were eligible to receive free or reduced-price lunches (Chicago Public Schools, 2020).

Like many other large urban school districts, CPS primarily employed punitive methods of student discipline in the past. In the 1980s and 1990s, the district implemented zero-tolerance policies mandating the use of suspensions and expulsions in response to student misconduct. These policies came under scrutiny at the federal, state, and local levels due to high suspension rates, especially among students of color (Stevens et al., 2015). They had important distributional consequences as students from the most vulnerable backgrounds – such as those living in poverty, those with disabilities, and those with a history of abuse or neglect – were more likely to be suspended (Sartain, Allensworth and Porter, 2015).

In the past decade, school districts across the country have started to recognize the potential adverse effects of such zero-tolerance policies on student outcomes and introduced alternative approaches in response to misconduct violations. In 2014, CPS announced a disciplinary policy reform plan called the Suspensions and Expulsion Reduction Plan (SERP),

with the goal of decreasing the number of out-of-school suspensions and expanding resources and training on school discipline to school staff across the district. This spurred various policy changes through the student code of conduct which included removing suspensions as a disciplinary response for a certain tier of infractions,[6] limiting the length of suspensions for substance use infractions, requiring district administrator approval for suspending students for certain behaviors, and most recently by removing in-school suspension for first-time lower-level infractions. These efforts are specifically expected to reduce inequities in suspension rates by race and other student characteristics (Sartain, Allensworth and Porter, 2015; Lai, n.d).

III.A Rollout of Restorative Practices Programs at CPS

In SY14, as a part of the SERP reform and as the district transitioned away from zero tolerance policies, CPS's Office of Social and Emotional Learning (OSEL) began to roll out district-wide restorative practices (RP) programs. This initiative was meant to not only give teachers clear guidance on alternative tools to suspension but also to improve the school environment itself. The district started by working with 22 high schools and 34 elementary schools in SY14 after CPS received a grant from the U.S. Department of Justice (DOJ) to help with their initial rollout of RP programming. By SY19, they expanded their RP programs to reach 279 schools, including 73 high schools.

The district offered different programs to support high schools in the adoption of restorative practices. These programs included RP Coaching, RP Leadership, and RP Peer Council.[7] Each of these programs is based on the same fundamental RP principles: community building, social and emotional learning, accountability, healing and reparation of harm, and restorative systems and mindsets.

The most intensive of these programs was RP Coaching, in which an RP coach trained administrators and designated individuals on a "School Climate Team" to model and implement restorative practices within their school. The school-based School Climate Team included one to two RP Leads who were responsible for training other staff and serving as a champion for RP throughout the building. The other members of the School Climate Team were to otherwise reflect the organizational composition of the school community, including a principal or assistant principal, dean or staff member responsible for disciplinary decision-making at the school, teachers representing all grade bands and subject areas, non-

[6] For grades three through twelve, out-of-school suspensions are now only permitted if a student's attendance endangers others, causes chronic/extreme interruption to others' participation in school, and prior interventions have been used. For students in kindergarten through grade two, central administration approval is required for any suspension.

[7] Appendix Table A1 presents summary statistics on the number of high schools by first RP type by school year. Some schools implemented a combination of multiple RP types in the same year.

teaching staff such as security officers and cafeteria workers, and family/community/student representatives.

The RP coaches were initially drawn from 15 different vendors with specialists who had expertise in restorative justice and how to adapt to different and dynamic school situations.[8] Typically, professional development is handled "in-house," or by existing staff, but this feature of bringing in outside experts was important because most staff were not otherwise trained in these approaches. Providing outside expertise facilitated adoption by school staff who did not otherwise feel equipped to engage with students in this way. Coaches came to schools and met with teachers, administrators, and other designated school staff two to three times each week throughout the academic year. This flexible model was designed to serve as ongoing professional development and meet schools' needs and abilities in developing a menu of restorative practices that was most appropriate for the context of their school and that could adapt to evolving situations. Once the DOJ funding ended in SY16, CPS reduced the number of vendors from which they drew and also reduced the frequency of coach engagement in schools to one day per week.[9] They also had to reduce the number of schools to which they could roll out RP programming.

The second program was RP Leadership, which entails a lighter touch intervention in schools. In RP Leadership, similar to RP Coaching, OSEL aimed to strengthen internal leadership capacity for building sustainable school-wide systems that foster community ties and address behavioral concerns. In these schools, OSEL typically focused on training a smaller number of school administrators for a much shorter amount of time. The third program, RP Peer Council, was a student-led process in which a small group of trained and designated students worked with referred students (who were involved in misconduct incidents or conflicts) to understand the impact of their actions on other individuals and school culture. Our evaluation focuses on understanding the impact of restorative practices in CPS high schools as a whole, although we also examine heterogeneity by program intensity to understand the differential role of implementation.

Schools were selected to receive restorative practices programs based on a variety of factors including a school's interest, a school's out-of-school suspension rate, a school's suspension rate for "priority" student groups, a school's climate indicators on the "My Voice My School" (MVMS) survey (now known as the CPS 5Essentials survey), school size, and input from those directly working with the schools (network specialists).[10]

[8] The longer that coaches stayed involved with a school, the more likely they would be incorporated as regular CPS school staff.

[9] As of SY22, half of the RP coaches are drawn from CPS staff, many of whom transitioned into CPS after working for the original RP vendors (with a particularly large share moving to CPS from one local vendor, Alternatives).

[10] "Priority" student groups have historically included students with Individualized Education Programs

IV Data Sources and Sample

Our analysis draws on four main sources of data: RP programming information from CPS's Office of Social and Emotional Learning (OSEL), CPS administrative data on students, CPS data on student responses to the MVMS survey, and Chicago Police Department (CPD) arrests data.

IV.A Data

Restorative Practices Programming Data. To identify the timing of treatment for students enrolled in a given school, we use programming data provided to us by CPS OSEL. These data include records on which schools first received restorative practices training in each school-year between SY14 and SY19 as well as the type of training received. A given school may have received multiple RP interventions. As such, in analyses that characterize differential impacts by RP program type, we assign schools to the first RP program type received.[11]

Student Administrative Data. We use CPS's student-level administrative data from SY09 to SY19 for information on student-level outcomes and demographics. The outcome variables include records of in-school and out-of-school suspensions, attendance records, GPA, and reading and math test score measures. The demographic information includes data on student race, gender, a proxy for economic disadvantage (whether the student is eligible for free or reduced-price lunch), housing status,[12] engagement with special education (IEP) or a 504 plan which indicates a physical and/or cognitive disability, and English learner status for those enrolled in CPS. Additionally, the data set includes information on student-level enrollment history, which we link to the programming data files from the OSEL to construct a student-level measure of treatment exposure.[13] We describe these data in more detail in Data Appendix C.

School Climate Data. Since 2011, CPS has administered annual surveys called "My Voice, My School" (MVMS) to understand the experiences of students in the school environment. To investigate the impact of restorative practices training on school climate, we examine student responses to this MVMS survey. The student survey is administered to students enrolled in grades six to twelve and comprises 21 constructs. We create a cli-

(IEPs) and Black students since these are the student groups suspended at the highest rates.

[11] If schools received multiple RP programs in the same initial year, we assigned them to the most intensive of the RP programs in which they participated (i.e., schools were assigned to RP Coaching if they began participating in RP Coaching and RP Leadership in the same school year).

[12] Homeless students are identified in CPS data as Students in Temporary Living Situations (STLS).

[13] CPS maintains a general student database in which each student is identified by a unique student ID. The distinct CPS administrative files are linked together by this ID.

mate index using data from student responses to eight of these constructs that speaks to a student's perceptions of school climate that may be directly affected by the introduction of RP. These constructs include Emotional Health, Student Classroom Behavior, Academic Personalism, Psychological Sense of School Membership, Personal Safety, School-Wide Future Orientation, School Safety, and Student-Teacher Trust (UChicago Impact, 2021). We use as placebo checks two constructs that ex-ante would not be expected to respond to school-based restorative practices (these constructs characterize parent supportiveness and community resources).

Police Arrest Data. Each of the data sources we have described are derived directly from CPS. Therefore, we draw on data from the Chicago Police Department (CPD) both to examine whether RP had a material effect on child behavior outside of the school context and to have an independent, externally constructed measure of child behavior. These data include individual-level arrest records from July 1, 2008 through June 30, 2019 and allow us to explore the effects of RP on juvenile arrests. The arrest data include information on the type (violent or non-violent offense), the location, and the time of arrest. We separately investigate the impact of restorative practices by arrest type and by whether arrests took place in school versus outside of school. We define "out of school" based on the location of the arrest occurring off of school grounds or the arrest occurring outside of school hours.[14]

Prior research has demonstrated that student arrests are associated with worse long-term outcomes, highlighting the importance of including this outcome measure in our analyses (Kirk and Sampson, 2013). The arrests-based analysis also allows us to explore whether any measured changes in school-based disciplinary outcomes (in our setting, suspensions) are driven by changes in teacher and administrator responses to misconduct rather than by changes in student behavior. To the extent this is the case, we would expect changes in arrests (and out-of-school arrests, in particular) to be muted in comparison to changes in disciplinary outcomes initiated by school staff.[15] The CPD and CPS data files are joined using probabilistic matching over a child's name, date of birth, gender, and home address.

IV.B Study Sample

Our analysis includes observations from students who were enrolled in any CPS traditional (district-run), contract, or charter high school between SY09 and SY19 for at least one day.[16] Table 1 presents average characteristics for students enrolled in the 184 CPS

[14]Specifically, in-school arrests are classified as incidents happening both inside the school location and reported between 7:00 AM and 6:59 PM during school days.

[15]It is important to note that the arrests data have no information about convictions, so included individuals may not have actually committed the criminal offenses for which they are arrested.

[16]As a robustness exercise, we also run our analysis restricting the sample to just district-run high schools. We discuss our findings in Section VI.

high schools in our sample in operation in the school year prior to the roll-out of restorative practices (SY13), separately for schools that did and did not receive any restorative practices programming at some point between SY14 and SY19.[17] This table shows that high schools that received RP training differed from never treated high schools in several ways at baseline. Treated high schools were significantly larger, with about twice as many students enrolled. Students in treated high schools had more absent days and lower GPAs at baseline, as well as more negative perceptions of their school climates.[18] Finally, treated high schools were also more likely to use suspensions as disciplinary tools. Though differences are not statistically significant at conventional levels, students who enrolled in subsequently treated schools had on average 38% more in-school suspension days (0.47 versus 0.34) and 24% more out-of-school suspension days (1.03 versus 0.83) than those enrolled in never-treated schools. As noted, the average differences we identify between subsequently treated and untreated high schools motivate our choice to employ a research design that relies on parallel trends-type (rather than strict exogeneity-based) identifying assumptions.

We focus our main analysis on high school students for two primary reasons. First, high school students are more likely than elementary school students to be arrested, both in school and out of school. For example, our analysis suggests that in SY13, 2% (6%) of high school students were arrested in (outside) CPS schools, compared to 0.3% (0.5%) of elementary school students. The low baseline rate of arrests in elementary schools poses a measurement challenge limiting our power to detect potential impacts on this margin and so to distinguish student behavioral responses from teacher-side responses to the introduction of RP. Second, student survey data on school climate, which permits us to investigate potential mechanisms driving estimated impacts on student outcomes, has limited elementary-school coverage.

V Research Design

In our benchmark specifications, we study the impact of the introduction of restorative practices on student disciplinary and academic outcomes, as well as measures of juvenile arrests. We also investigate how student perceptions of school climate respond to the roll-out of restorative practices. The criteria used to allocate RP programming motivates our difference-in-differences research design. Specifically, since schools receiving RP programming are likely to differ on various dimensions when compared to schools not receiving RP

[17] Appendix Tables A2 and A3 present average characteristics by demographic group and based on alternative sample partitions. Appendix Table A4 presents average characteristics for students enrolled in CPS elementary schools in our sample in SY13, separately for students in schools that did and did not receive any restorative practices programming at some point between SY14 and SY19.

[18] To ensure that our attendance measure is not mechanically correlated with our measure of out-of-school suspensions, we subtract the number of out-of-school suspension days from total number of absences. In-school suspension is not considered an absence because the student is still in a supervised setting inside their school.

programming, our research design relies on a weaker conditional exogeneity assumption that requires that expected *changes* over time in outcomes absent treatment are independent of RP programming assignment.

To identify treatment effects associated with exposure to restorative practices, we rely on variation in exposure induced by the rollout of RP over time and across schools. Since student enrollment choices may respond endogenously to RP exposure, we identify student-level treatment exposure based on the first high school that each student attended within the CPS system, as well as the year and grade level in which that student enrolled in CPS.[19] To guide thinking, if student i attended high school g from SY10 to SY12, and then moved to high school g', the student's treatment exposure remains a function of the timing of RP rollout in school g, regardless of whether the first intervention in school g occurred before or after the student had transferred. The subsequent analysis includes one observation per year per student for every student who was enrolled for at least one day in any CPS high school in the corresponding year, according to the enrollment history files.[20,21]

Our identification assumption is that students enrolling in schools that did and did not adopt restorative practices over a given period would have exhibited parallel trends in relevant outcomes in the absence of the rollout of the restorative practices treatment. An extensive recent literature has highlighted that estimators derived from standard two-way fixed effects models employed to identify treatment effects in settings with multiple treated groups and staggered rollout of treatment are unbiased only if treatment effects are homogeneous across time and group (Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfoeuille, 2020; Sun and Abraham, 2020). In practice, however, there are a number of reasons to hypothesize that the effect of exposure to restorative practices may vary with intensity (i.e., number of years) of exposure as well as the timing of introduction. First, student outcomes may be a function of cumulative exposure to restorative practices to the extent that behavioral changes take time to manifest. Second, teachers' disciplinary practices, and school climate more generally, may evolve over time as the core principles of restorative practices become more ingrained. Third, the quality and refinement of RP programming over time may generate treatment effect heterogeneity as a function of the timing of its introduction. This anticipated treatment effect heterogeneity (which is ultimately borne out in the data)

[19] Since enrollment records are unavailable prior to SY09, we assign students enrolled in CPS prior to SY09 to schools based on their SY09 enrollment record.

[20] We exclude the following observations from this sample: students who have progressed to grade levels not offered by their initial schools, students past their expected school exit year, and any observations beyond our event study window (-5 to +5 years since treatment for all outcomes other than school climate, -3 to +4 years since treatment for our school climate outcome given a lack of available MVMS survey data at the start and end of our sample period) from students assigned to treatment schools.

[21] We follow an analogous approach when analyzing outcomes for students in elementary schools.

implies that standard two-way fixed effects models are inappropriate for our study setting. The resultant bias arises from the fact that standard two-way fixed effects models rely on already-treated groups when constructing counterfactuals; to the extent that changes in outcomes in these already-treated groups are themselves partly driven by the dynamic effects of the treatment, this comparison introduces bias. As shown in Sun and Abraham (2020), even event study models that separately estimate the effects of treatment as a function of treatment timing will be biased in the presence of such treatment effect heterogeneity. The fact that a sizable share of CPS high schools is ultimately treated indicates that accounting for treatment effect heterogeneity is particularly important in our study setting.

To test our identifying assumptions and estimate the causal effect of restorative practices in the presence of heterogeneous treatment effects, we rely on an estimator derived in de Chaisemartin and D'Haultfoeuille (2020), which is designed to produce unbiased estimates of the average effect of treatment on the treated (both averaged across post-treatment periods and separately by treatment timing) when such heterogeneity is present. This estimator uses only not-yet-treated groups (in our study setting, students assigned to not-yet-treated schools) to predict counterfactual outcomes and so ensures that treatment effect estimates are not contaminated by treatment-induced changes in outcomes in already-treated groups.

To formally characterize the de Chaisemartin and D'Haultfoeuille (2020) estimator in the context of our study setting, we define $D_{i,g,t}$ as an indicator for restorative practices exposure of student i with assigned school g in school year t. We classify each school as exposed to the restorative practices treatment in all years after its introduction; in practice, we cannot measure whether restorative practices continued to be employed in subsequent years.[22] Following the notation from the authors' derivation, we define $N_{g,t}$ as the number of students assigned to school g in school year t and we define $N_{d,d',t} = \sum_{g:D_{g,t}=d,D_{g,t-1}=d'} N_{g,t}$ as the total number of students assigned to schools in school year t that had treatment value d' in school year $t-1$ and treatment value d in school year t (where the treatment value is equal to zero if the school had not introduced RP and equal to one if the school had introduced RP). Next, we define:

$$(1) \quad DID_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}(Y_{g,t} - Y_{g,t-1})$$

This expression returns a weighted average of the difference between the change in outcomes between school year $t-1$ and t in schools first treated in school year t and the change in outcomes between $t-1$ and t in schools untreated through school year t. As

[22] To the extent that a subset of schools transitioned away from restorative practices, our treatment effect estimates will consequently represent lower bounds on the true causal impact of persistent RP exposure.

shown in de Chaisemartin and D'Haultfoeuille (2020), we can then take a weighted average of $DID_{+,t}$ across all school years from $t = 2$ to $t = T$ (where T is the final school year in the study sample) to produce an unbiased estimator of the average treatment effect in the first post-treatment school year of all schools that become treated during the sample period.[23] Specifically, the weighted average is constructed as follows:

$$DID_M = \sum_{t=2}^{T} \left(\frac{N_{1,0,t}}{N_S} DID_{+,t}\right) \tag{2}$$

where N_S is the total number of students in the year that their assigned school is first treated. Finally, we employ this same approach to construct treatment effect estimates specific to the number of school years since initial treatment exposure and, alternatively, as a function of the number of school years until initial exposure. These latter placebo estimates can then be used to evaluate the parallel trends assumption, as in the standard event study framework.

Turning back to the study setting, one key challenge is that we are interested in analyzing changes for a wide range of outcomes in response to the introduction of restorative practices. Since the parallel trends assumption must be evaluated for each outcome of interest, we present event study plots for all outcomes subsequently analyzed in our main tables. Following the notation used above, $Y_{i,g,t}$ is the outcome of interest for student i who was first enrolled in school g and is being observed in school year t, and $Y_{g,t}$ is the corresponding average outcome value in school year t for students assigned to school g. RP programming was introduced across grade levels within adopting schools. Consequently, $D_{g,t,l}$, our treatment measure where l corresponds to the number of years since treatment, is an indicator defined by whether restorative practices were introduced in school g exactly l years after school year t (or $|l|$ years before for negative-valued l).

Across analyses, our benchmark models also include the following student-level covariates: age fixed effects,[24] cohort fixed effects,[25] gender fixed effects, race fixed effects (students who identify as Asian, Black, Latine, White, or other races), an indicator for unhoused, an indicator for whether the student is enrolled as an English Learner, indicators for student disability classification from engagement with either special education (IEP) or a 504 plan which indicates a physical and/or cognitive disability, and an indicator for whether the student is eligible to receive free or reduced-price lunch.[26] In practice, the inclusion of these

[23] The subscript + indicates the change from untreated to treated status.

[24] Age is defined as the student's age by June 20 of the last calendar year of the school year (the last possible end date for a school year).

[25] Cohort defines a set of grade levels and school years corresponding to the same set of students in the absence of entry/exit or grade retention (i.e., one cohort includes ninth grade students in SY11, tenth grade students in SY12, etc.).

[26] In specifications that employ absent days as the outcome of interest, we also include yearly total "member

covariates improves the precision of estimates in some instances, but does not alter the basic pattern of findings nor our conclusions regarding the validity of the parallel trends assumption that underpins the research design.[27] Across analyses, to account for the school-level nature of treatment assignment, we cluster standard errors at the level of the school in which each student first enrolled.

The event study plots for high-school student outcomes, presented in Figures 1 and 2, provide support for the parallel trends assumption with respect to key outcomes of interest.[28] In subsequent analyses, we combine estimates of the instantaneous and dynamic effects of restorative practices exposure to produce a single estimate of the causal effect of treatment on the treated for each outcome of interest. To do so, we construct the following estimator of the average cumulative effect of restorative practices over $k+1$ treatment periods (where k is set to 5 to avoid small cell sizes):[29]

$$\hat{\delta}_{+,0:k} = \sum_{l=0}^{k} \omega_{+,k,l} DID_{M,l} \tag{3}$$

Here, $DID_{M,l}$ is defined analogously to DID_M and captures the weighted average effect of treatment l periods after initial treatment exposure. $\omega_{+,k,l}$, the weight assigned to the treatment effect l periods after initial treatment exposure, is defined as $\frac{N_l^1}{\sum_{l=0}^{k} N_l^1}$, where N_l^1 is the number of students in the sample l school years after initial treatment exposure by the end of the study period (year T, corresponding to SY19).

VI Main Results

We seek to understand the role that school behavioral policies may play in shaping child behavior and perceptions. Specifically, we analyze the shift from more punitive practices to more restorative practices in response to perceived student misconduct and examine how children's behavioral outcomes, educational outcomes, and perceptions of school climate changed.

Changing behavior inside of school. First, we examine the impact of the introduction of restorative practices on in-school behavioral outcomes. Figure 1 shows an event study plot that is indicative of growing declines in out-of-school suspensions in the years after initial treatment exposure. Aggregating instantaneous and dynamic estimates, we identify a

days" as a control. Member days represents the sum of the number of days that a student was present in school and the number of days that the student was absent from school.

[27] To incorporate covariates, differences in counterfactual outcomes across periods are allowed to vary linearly based on changes in group-level average covariate values.

[28] The event study plots for elementary-school students are presented in Appendix Figures A1 and A2.

[29] k is set to 4 for the school climate outcome because data are available for one fewer year after the introduction of RP.

significant decrease in out-of-school suspensions of 0.17 days, or 18 percent (Table 2, column 1). By contrast, estimated impacts on in-school suspensions and days absent are negative but statistically indistinguishable from zero (Figure 1; Table 2, columns 3, 4, and 5). Taken together, these findings suggest that students are receiving more in-school instruction time, on average.[30] The measured decline in suspensions serves as evidence of a "first stage" – RP changed the behavior of teachers and/or students.

Changing behavior outside of school. We are interested in understanding whether being exposed to restorative practices affects conflict resolution regardless of location and separate from structured or guided intervention. To examine whether being exposed to restorative practices is changing student behavior rather than simply changing how adults in schools respond to student behavior, we draw on arrest data from Chicago Police Department (CPD). Police officers serving outside of schools are not under the same authority as teachers and operate independently from school policies and practices. Consequently, arrest records can be used to produce an independent measure of perceived student behavior.

In Figure 1, Panel D, we show an event-study plot for number of arrests, which exhibits a relatively flat pre-trend followed by a decline in arrests that increases in magnitude with time since the introduction of restorative practices. The estimated aggregate impact is an average decrease of 0.024 arrests, which represents a 19 percent decline relative to the baseline mean (Figure 1; Table 3, column 1).[31,32]

While the estimated decline in arrests in response to the introduction of restorative practices is consistent with improved student behavior, school staff are tasked with referring students to law enforcement when they feel that they need an external disciplinary authority to intervene on matters that occur at school. Consequently, decreases in juvenile arrests could still reflect the fact that adults in schools are induced to reduce overall punitiveness in response to the introduction of RP. To distinguish between alternative explanations for the aggregate decline in student arrests, we next examine whether reductions in arrests were driven entirely by arrests made on school property and during school hours, or whether we also identify declines in arrests outside of school grounds or school hours, which would suggest that changes are not solely driven by school staff referral behavior but rather by genuine changes in student behaviors and approaches to conflict resolution. In Table 3, columns

[30] We find parallel evidence of declining out-of-school suspension days for elementary school students with imprecisely estimated impacts on in-school suspension days and absent days for these grade levels (Appendix Table A5, Appendix Figure A1).

[31] We also estimate a decline in the likelihood of being arrested for elementary school students (Appendix Table A6, column 1), though the coefficient is small in magnitude and not statistically distinguishable from zero. The baseline arrest rate for elementary school students is quite low, so there is less scope for large impacts on this margin.

[32] Appendix Table A7 presents the estimated impact of RP on high school students' binary arrest outcomes.

2 and 3, we provide evidence that aggregate arrest declines reflect decreases in in-school and, separately, out-of-school arrests (by 34.6 percent and 14.7 percent, respectively). These findings provide evidence in support of the hypothesis that student behavior is responding to the introduction of restorative practices.

A broader question is whether a restorative justice approach to conflict can decrease violence.[33] To explore this question, we examined changes in arrests separately for violent and non-violent offenses. We see declines in arrests for both types of offenses: a 17.9 percent reduction in the number of violent arrests and a 20 percent reduction in the number of non-violent arrests (Table 3, columns 4 and 5), suggesting that the introduction of RP also led to a decrease in violence.

Changing school climate. We saw that the introduction of restorative practices resulted in a decrease in out-of-school suspensions (Table 2, columns 1 and 2), and the declines in out-of-school arrests suggest that this effect is not simply the mechanical result of teachers being under explicit instruction not to suspend students. As such, estimated RP impacts likely reflect some combination of changes in adult behavior (for instance, how they interact with and understand students) and student behavior (for example, how students respond to conflict or to feeling more understood by adults in school and their peers). Consistent with the hypothesis that restorative practices engender genuine changes in staff and student attitudes and behaviors, we find significant improvements in student-reported measures of school climate (Table 4). Specifically, we identify a 0.042 standard deviation improvement in perceived school climate. This aggregate impact is driven by particularly large increases in students' perceptions of their peers' classroom behavior, in their psychological sense of school membership, and school safety (Table A8). We do not, however, see corresponding changes in our placebo measures – student perceptions of parent supportiveness or human and social resources available in the community – which we would not expect to be affected by a school-based introduction of RP.

Examining student learning. Despite these improvements in school climate, we do not see any corresponding evidence of improvements in academic performance as measured by GPA, reading test scores, or math test scores (Table 4, columns 2, 3, and 4).[34] However, in contrast to behavioral outcome measures that are rarely missing for enrolled students, we note that the rate of test score missingness exceeds 18% within tested grade levels during the study period. RP adoption leads to increased classroom time and has the potential to affect missingness rates. In Appendix Figure A3, we show that test score missingness indeed

[33]Such reforms are being experimented with within criminally-accused situations. Results from our setting could inform practices in contexts separate from schools.

[34]We also find null effects among elementary school students (Appendix Table A9, Appendix Figure A2).

falls by 2 to 2.5 percentage points in response to RP adoption, which may introduce bias to the extent missingness is non-random. To probe the extent of such bias, we impute test score values that are missing using alternative measures of student performance.

To impute test scores, we rely on a gradient boosting method, which generates predictions based on constituent "trees." Each tree in the forest is "grown" by sequentially partitioning the data to maximize prediction accuracy, with all observations landing in a given partition being assigned a uniform prediction. Each tree develops its own prediction model, and the gradient boosting fits the sequential collection of regression trees such that each tree is optimized to learn from the prediction error of the trees before it (Ke et al., 2017). In practice, we predict contemporaneous test scores using the student's lagged test score, contemporaneous GPA, and indicators for whether these outcomes are missing. To capture any differential predictive power across standardized test instruments, we also include year-by-grade indicators as categorical predictors.[35] Imputation-based results (shown in Table 4, columns 5 and 6) suggest that differential test score missingness does lead to downward bias, although the magnitude of such bias appears limited. We cannot reject null test score impacts after accounting for differential test-taking rates as a function of RP status.

A common concern is that reducing suspensions of students who engage in undesirable behaviors keeps these students in the classroom and they may then disrupt the learning of their peers. While we do not identify any improvements in academic performance in response to the introduction of RP, the shift away from punitive, incapacitation-focused disciplinary responses also does not seem to have been detrimental to the learning outcomes of the broader student body, on average. This basic conclusion is reinforced by student self-reports indicative of improved student classroom behavior (Table A8). Nonetheless, in Section VII, we directly test for the presence of disruption effects by exploiting variation in student exposure to classmates at high risk of suspension (who experience differential declines in out-of-school suspension days in response to the introduction of RP). We then probe the extent to which treatment effects vary by student characteristics in order to further unpack our average findings.

Implementation matters. Before summarizing the sensitivity testing we undertake to probe the robustness of our findings, we briefly assess how program intensity affects our research conclusions. Restorative practices comprise a wide range of implementation ap-

[35] We used the `scikit-learn` implementation of gradient boosted trees and ran the algorithm for 100 iterations with a learning rate of 0.1. To avoid overfitting, we set the L2 regularization parameter to 1.0, enforced a max depth of 10 on each tree, and requested that each leaf contain a minimum of 100 observations. We chose these parameters after a "tuning" process that canvassed multiple parameter combinations to eventually select the combination with the best prediction accuracy.

proaches, which can make it hard to replicate and scale successful models. To understand what specific set of practices was most effective, we explore differential impacts for the different models implemented in high schools: RP Coaching, RP Leadership, and RP Peer Council.

RP Coaching was the most intensive of the programs and involves an RP coach who trains administrators and designated staff to model and implement restorative practices and then meets regularly with staff throughout the school year. RP Leadership was a lighter touch intervention in which a smaller number of school administrators are trained for a much shorter amount of time. RP Peer Council was a student-led process in which a small group of student members of the peer council work with students who were involved in misconduct incidents or conflicts.

In Appendix Tables A10 and A11, we show suggestive evidence that RP impacts on behavioral outcomes are mainly driven by the RP Coaching approach, the most intensive of the RP practices. It is important to note, however, that the relative infrequency with which schools have participated in the RP Leadership and RP Peer Council programs means that those program-specific treatment effect estimates are generally imprecise.[36]

Attrition. The decline in test score missingness associated with RP adoption that we have documented is likely driven in part by the increased time that students spend in the classroom, but may also reflect a decline in school exit. We next test explicitly for differential attrition in order to understand the potential for selection bias more broadly (recall that a student who does not attend any CPS school in a given year is absent from our study sample). To do so, we construct an artificial panel in which we include one observation for each student in each grade level between nine and twelve under the assumption that students progressed one grade level each year. For those student-grade observations that do not appear in our study sample due to student attrition, we code an attrition indicator variable equal to one. In our setting, attrition may arise from student transfers to private schools, movement to districts outside of CPS, or student dropout. In regression analyses that parallel our benchmark models but employ this attrition indicator as the dependent variable, we find that attrition declines by 1-2 percentage points in response to the introduction of RP (depending on whether covariates are included in the model). In Appendix Figure A4, we first present an event study plot characterizing attrition rates as a function of treatment timing. We next show graphically that attritors in schools that do and do not implement RP are more likely than their non-attriting classmates to be suspended in their first year in a CPS high school

[36]Moreover, while schools that implemented RP Leadership did not subsequently implement RP Coaching, interpretation of RP Peer Council treatment effects is complicated by the fact that several schools implementing RP Peer Council subsequently implemented RP Coaching as well.

(the availability of first year data is not affected by subsequent attrition). Given that the characteristics of attritors do not appear to vary by school RP status, we conclude that the RP-induced reduction in attrition we identify would, if anything, be expected to attenuate the estimated beneficial impacts of RP on suspension days and other correlated outcomes, including arrests. In the context of our null average findings on academic outcomes, it is similarly possible that the differential selection we identify would mask any underlying achievements gains. Although interpreting attrition is challenging given that we cannot distinguish transfers from dropout, a reduction in attrition is consistent with the notion that students in RP-implementing schools feel a greater sense of belonging and are consequently less likely to exit.

Additional Sensitivity Analysis. We investigate the sensitivity of results to a range of alternative empirical approaches and specifications. We confirm that results remain robust across these alternative modelling choices.

Standard difference-in-differences empirical approach. Instead of using the de Chaisemartin and D'Haultfoeuille (2020) estimator, we employ a standard difference-in-differences design. The results remain qualitatively similar to the effects estimated in our benchmark specifications, with a notably larger estimated decline in absent days driven by the differential pre-trends apparent for this outcome (Appendix Tables A12 and A13, Panel A).[37]

Excluding charter and contract schools. For our main specifications, we include all observations for students who were enrolled in district-run, charter, or contract schools in a given school year. To check the sensitivity of our results, we restrict the sample to students who remained in traditional district-run schools and so exclude all observations for students who ever attended a charter or contract school in a given school year. The results remain largely unchanged (Appendix Tables A12 and A13, Panel B).

Alternative specifications. We verify that results are not sensitive to the exclusion of covariates by estimating models that include only age and cohort fixed effects. We find qualitatively similar results, although the modest decline in GPA is now significant at the 10% level (Appendix Tables A12 and A13, Panel C).

VII Mechanisms: Disruption Effects

While null impacts on academic performance outcomes suggest that the disruptive effects of students who would be suspended in the absence of RP are likely limited, the possibility of such disruption effects is a key concern among those who advocate for the

[37] Appendix Figures A5 and A6 present the event studies around the introduction of RP using a standard difference-in-differences approach.

status quo of more punitive disciplinary practices. This concern is premised on the notion that those students who were being suspended under the status-quo system were those most likely to disrupt student learning. If, after RP, these students were now less likely to be removed from the classroom via a suspension (we see a decline in OSS in Figure 1 and Table 2), their classmates who were themselves at low risk of suspension could be differentially harmed by the introduction of RP. Said differently, while students at risk of suspension may benefit directly from the introduction of RP through increased engagement and an increase in instructional time, it is possible that those who were already suspended at low rates at baseline (and so mechanically stand to benefit less from RP adoption on this margin) may be harmed academically. To rigorously test for the presence of such disruption effects, we exploit variation in student-level exposure to potentially disruptive peers. Specifically, if students unlikely to be punitively disciplined at baseline experience test score declines as a result of disruption effects, the magnitude of such declines should be growing in the decrease in out-of-school suspension days faced by the student's classmates.

To generate variation in the RP-induced decline in out-of-school suspension days faced by each student's classmates, we employ a random forest algorithm. We use this algorithm to predict who may have been more likely to have been affected by the policy shift, i.e. those more likely to have been suspended under the prior, more punitive regime. Specifically, we predict high school suspension days based on a rich set of eighth grade characteristics that include student race, gender, number of arrests, attendance, GPA, and out-of school-suspension days. In addition, we allow predictions to vary with a number of characteristics measured contemporaneously (in high school): free or reduced-price lunch eligibility, English-language learner status, and unhoused status.

In our setting, the random forest algorithm offers a data-driven approach to identifying the optimal prediction model, allowing for arbitrary interactions between included covariates and relaxing the parametric assumptions imposed in standard linear regression models. Here, each tree in the forest is "grown" using a predetermined fraction of the available predictor variables, and the data used to "grow" each tree is sampled with replacement from the original data set. This bootstrap aggregation ("bagging") strategy aims to reduce the tendency for any given tree to have high variance on its own (i.e., to learn a prediction model that generalizes poorly).[38] We employ only data from SY13 and earlier to construct the random forest.[39] This reliance on measures collected prior to the introduction of RP to define

[38] See Breiman (2001) for further details on the bagging involved in the random forest algorithm.

[39] The larger set of potential predictors here (as compared to when imputing test scores) motivates the choice to employ a random forest algorithm. The random forest was implemented via the algorithm developed in the open-source `H2O.ai` platform. All hyperparameters were kept at their default values in the `H2O.ai` implementation: the number of trees is set to 50, the maximum depth of a tree to 20, and the number of

the prediction model ensures that predictions are not influenced by the effects that RP may itself have on the link between student characteristics and high school student outcomes. Moreover, by focusing only on the prediction based on baseline suspension rates (as opposed to the effect of RP on suspension rates), we rely on the testable hypothesis that RP-induced suspension declines will be largest where the predicted baseline suspension rates are highest.

To classify students (both prior and subsequent to SY13) based on classmates' predicted baseline suspension rates, we construct student-level predicted high school suspension days using the random forest algorithm described above.[40] We then construct predicted suspension day averages at the school-by-cohort level, and we partition school-by-cohort cells into above- and below-median average predicted suspension day groups within a given cohort. We refer to these groups as "above-median" and "below-median" for brevity. Finally, we re-estimate our benchmark regression models separately for students in above- versus below-median predicted suspension day cells.

The results, presented in Table 5, column 1, validate the use of predicted suspension days to generate heterogeneity in RP-induced suspension day declines. Students in above-median cells experienced a 0.24 day decline in out-of-school suspension days in response to adoption (45% larger than our full sample estimate) compared to students in below-median cohort-by-school cells, who experienced a 0.096 day decline in out-of-school suspension days. In columns 2 and 3, we see evidence of heterogeneity in test score impacts that (though imprecise) is consistent with there being direct returns to increased engagement and instructional time.[41] While below-median students experienced negative estimated changes in math and reading scores, above-median students experienced estimated gains in response to RP adoption.

Our proposed test for disruption effects requires that we identify students who vary in their exposure to potential disruption (i.e., declines in classmate suspension rates) but who do not themselves experience differential changes in suspension days. To implement this test, we focus on students who themselves have below-median predicted suspension days (with median values again constructed within cohort). We show in Table 5, column 4 that these students experience small and statistically insignificant changes in out-of-school suspension days in response to RP adoption (point estimates for students in below- and above-median predicted suspension day cells are -0.005 and -0.020, respectively).[42]

features for each tree to split on equals the number of predictors divided by 3.

[40] For observations corresponding to SY14 and later, we use the random forest algorithm results (based on pre-period data) and student characteristics to predict high school suspension days.

[41] This finding may also be explained by the greater scope for academic gains in those schools that are most disadvantaged at baseline in terms of academic outcomes, school climate and disciplinary challenges. Table A14 shows that students in above-median cohort-by-school predicted suspension cells indeed started with worse baseline outcomes than those in below-median cells.

[42] An alternative approach would be to compare all students in below- versus above-median predicted sus-

Having identified a set of students who themselves experience no significant change in suspension days in response to RP adoption (and so would not be expected to experience test score gains through direct increases in instructional time or engagement), we can test directly for disruption effects by examining whether those with high predicted classmate suspension rates experience larger test score declines in response to RP adoption. As shown in columns 5 and 6 of Table 5, we do not find evidence of heterogeneous test score impacts consistent with disruption effects. For students with below-median predicted classmate suspension rates, we identify an insignificant 0.056 SD decline in math test scores and an insignificant 0.031 SD decline in reading test scores. For students with above-median predicted classmate suspension rates, we identify an insignificant 0.031 SD increase in math test scores and an insignificant 0.007 SD decrease in reading test scores.[43] These findings represent a rejection of the disruption hypothesis. Though the imprecision associated with these estimates suggests that they should be interpreted with caution, any test score declines not explained by disruption effects may alternatively be driven by changes in the nature of teaching practices or classroom time usage that result from the introduction of RP.[44]

VIII Treatment Effect Heterogeneity by Student Characteristics

To understand the distributional implications of the average impacts we estimate, we consider treatment effect heterogeneity based on student characteristics: English-language proficiency, grade level, disability (either from engagement with an IEP plan or a 504 plan), race, and gender. Our analysis emphasizes differential impacts by student race and gender, which are two of the strongest observable predictors of baseline suspension and arrest patterns. For each source of heterogeneity analyzed, we conduct subsample-specific analyses and contrast treatment effect estimates (i.e., to investigate heterogeneity by English learner status, we separately estimate benchmark regression models using the subsample of English language learners and the subsample of native English speakers).

Heterogeneity by English learner status, grade level, and disability. Examining heterogeneity by English learner status, we find that reductions in out-of-school suspensions are concentrated among native English speakers (Appendix Table A15), suggesting that RP implementation, which can be nuanced and requires clear communication between parties,

pension day cells while conditioning on own predicted suspension days. In practice, however, we find that students in above-median predicted suspension day cells experience larger declines in suspension days in response to RP adoption, conditional on own predicted suspension days. This finding may be explained by the fact that students who are themselves at risk of suspension are more likely to be suspended when surrounded by other high-suspension propensity students due to peer effects.

[43] Appendix Figures A7 and A8 present event studies for estimated out-of-school suspension days and test score outcomes by classmates' predicted suspension rates.

[44] Classroom teachers may proactively dedicate instructional time to community-building activities, including in-class circles, designed to increase buy-in among students.

may have been better translated to those who were fluent in the instructional medium, which would have been English. Alternatively, these patterns may be driven by higher rates of suspension among native English speakers at baseline.[45]

We also find larger absolute declines in the number of out-of-school suspension days and arrests for 9th and 10th graders, who are suspended and arrested more frequently at baseline, as compared to 11th and 12th graders (Appendix Table A17).[46] This is consistent with there being more room to adopt new practices and norms when one is newer to a setting (as 9th and 10th graders are to high schools) as opposed to once students have acclimated to a certain culture, which 11th and 12th graders would have been more likely to have done.[47]

With respect to engagement with either special education (IEP) or a 504 plan which indicates a physical and/or cognitive disability, we find that declines in out-of-school suspensions do not vary significantly with disability status, while estimated declines in arrests are notably larger for students with disabilities (Appendix Table A19).[48]

Heterogeneity by race and gender. Student race and gender are key predictors of baseline exposure to punitive disciplinary practices, and we find evidence of stark heterogeneity in RP responses as a function of these same characteristics. We begin by examining changes in out-of-school suspensions in response to RP adoption, and we find that the aggregate reductions in the number of out-of-school suspensions we estimate are driven by declines in out-of-school suspensions among Black male and female students, who experience declines of 0.384 and 0.325 suspension days, respectively (Table 6, column 1).[49] In Table 6 (column 3), we show that Black students similarly experience the largest absolute reductions in arrests (with estimated declines of 0.079 and 0.015 arrests for Black male and female students, respectively). While Black students are most frequently suspended and arrested at baseline, these large absolute declines suggest that they may differentially benefit from the introduction of restorative practices on other dimensions as well. Indeed, we see a significant decline in absent days among Black males (1.66 days, or 7.9%), above and beyond the identified reduction in out-of-school suspension days.

Turning to academic outcomes, this increase in instruction time for Black males is associated with significant math test score gains (more muted effects of educational pro-

[45] Appendix Table A16 presents the school climate and learning outcomes by English learner status.

[46] Appendix Table A18 presents the school climate and learning outcomes for 9th and 10th graders and, separately, for 11th and 12th graders.

[47] Alternatively, this may be explained by differences in baseline incidence levels and/or differences in enrollment persistence across grade levels.

[48] Appendix Table A20 presents the school climate and learning outcomes by disability status.

[49] Interestingly, the suspension day declines for Black students exceed the estimated decline (shown in Table 5) for students explicitly identified as being at high risk of suspension at baseline. This may reflect the salience of race as a driver of teacher responses to the introduction of RP or may reflect Black student behavior being particularly responsive to its introduction.

grams/policies on reading scores are consistent with prior work; see, for instance, the related discussion in Fryer, 2014). For Black female students, we identify particularly large improvements in self-reported school climate along with positive but more muted (and statistically insignificant) math and reading test score responses. In contrast, we find that Latino male students experience marginally significant test score declines (test score impacts are positive but insignificant for white males and negative but smaller in magnitude and insignificant for Latina females and white females). In Appendix Table A21, we present estimates based on imputed test scores, which are qualitatively similar but more muted for Latino males (and no longer significant at conventional levels).

One explanation for the differential gains experienced by Black male students is that they may be concentrated in those schools that employ RP *most* effectively, while Latine students may be concentrated in those schools that employ RP *least* effectively. To probe this possibility, we investigate heterogeneity by the school-level Black student enrollment share. Specifically, in Table 7, we re-estimate our benchmark model separately for the subset of students enrolled in schools with below-median and above-median Black-student-enrollment shares (relative to the median student's school-level share in SY13). In this analysis, schools are assigned to above- versus below-median bins based on their Black-student-enrollment share in SY13 or the first subsequent year in which the school appears in the study sample. We examine treatment effect heterogeneity with respect to out-of-school suspensions and student test scores in the full sample of students, as well as separately by race for the two largest racial groups in the sample (Black and Latine).

Table 7, Panel A, column 1 indicates that RP-induced declines in out-of-school suspensions are concentrated in schools with above-median Black-student-enrollment shares. However, Table 7, Panels B and C, which break down overall impacts by student race, reveal that effects on out-of-school suspensions are fairly uniform within each racial group for students in below- versus above-median Black student enrollment share schools. These findings are consistent with RP implementation being relatively homogeneous across schools, with Black students simply benefiting most in terms of reduced exposure to punitive discipline (point estimates for Latine students remain negative but statistically indistinguishable from zero in both below- and above-median schools).

Turning to student test scores, estimates are generally imprecise but we do see suggestive evidence that math test score impacts are more positive in schools with above-median Black-student-enrollment shares.[50] Although we are underpowered to identify whether these gaps are meaningful and to understand what explains them if so, our findings may reflect differ-

[50] Appendix Figures A9, A10, and A11 present the associated event study plots for all students, Black students and Latine students, respectively.

ences in how effectively teachers are able to adapt classroom practices to incorporate lessons from the RP curriculum, or may reflect the fact that the scope for academic gains is greatest where discipline-related disruptions were initially the most frequent, academic performance was initially lowest, and school climate perceptions were initially the most negative.[51]

IX Conclusion

Historically, parents have sent their children to school with an implicit trust that the policies and practices of a school, if implemented properly, would necessarily result in the best outcomes, not only for their children but also for society. School officials themselves, however, struggle with the decision as to which policies are optimal, particularly when establishing safety and disciplinary systems. Schools tend to be risk-averse, and the inherently "safe" option is to have no tolerance for any breaches of what is considered to be appropriate conduct. On the other hand, by enforcing an overly retributive system, schools may be inadvertently cultivating a less tolerant society and exacerbating already stark disparities for students from disadvantaged backgrounds. The lack of clarity regarding the costs and benefits of a more or less punitive system necessitates a rigorous evaluation of different school policies and practices that are implemented with the intention of improving behavior and increasing safety of the school.

We study the causal impact of the rollout of restorative practices in Chicago Public Schools. We use cross-school variation in the timing of the introduction of RP to understand how adoption of a restorative approach affects students' behavioral and academic outcomes, their perceptions of school climate, and their involvement with the criminal legal system. Our evidence suggests that the introduction of RP in CPS high schools reduced the number of out-of-school suspension days by 18 percent and reduced the number of student arrests by 19 percent. Correspondingly, we find significant improvements in perceived school climate in response to the introduction of RP, suggestive of genuine changes in underlying student behaviors and attitudes. We identify sizable declines in both in-school and out-of-school arrests, which further confirm that the changes in disciplinary outcomes we estimate do not simply capture changes in how teachers and school administrators respond to behavioral challenges. We also estimate meaningful decreases in arrests for both violent and non-violent offenses, suggesting that RP can be linked with declines in violence. We do not find any evidence that RP significantly impacts student grades or test scores in the aggregate, and the results we present are inconsistent with RP-induced disruption effects that represent a common concern among practitioners.

[51] At baseline, schools with higher shares of Black student enrollment have on average lower student GPA, lower perceived climate scores, and higher numbers of suspension days and arrests than schools with lower shares of Black student enrollment (see Appendix Table A22).

Turning to treatment effect heterogeneity, we find that absolute declines in the likelihood of out-of-school suspensions and arrests are largest among Black students, who face the highest suspension and arrest rates and have the most negative perceptions of school climate at baseline. The reductions in exposure to punitive measures are particularly beneficial for Black males, who attend school more frequently after the introduction of RP and experience significant math test score gains. In addition to reflecting the returns to increased instructional time, our findings are consistent with the notion that changes in responses to perceived behavior and improvements in climate may lead to academic gains where baseline punitiveness and disengagement is most prevalent. Regardless, our findings indicate that RP interventions like those we evaluate have the potential to meaningfully impact those students most exposed to punitive disciplinary practices at baseline. While some practitioners may be concerned that RP would benefit students who would otherwise be exposed to punitive discipline while harming their classmates by engendering more permissive behavioral norms, our results highlight that no such tradeoff exists. Future research should examine the longer-term implications of changes in disciplinary practices with regards to high school completion, post-secondary enrollment and future criminal legal system involvement.

School disciplinary policies may reach beyond the creation of conditions for learning in the short term. They may also send signals to children about optimal ways to behave and how society should ideally work (Parsons, 1959; Dreeben, 1967; Bowles and Gintis, 1976). When school districts rely on primarily punitive responses to resolve minor conflicts, children may infer that the optimal approach to undesirable situations is one of retribution. However, if a school district instead emphasizes a reparative or restorative approach to addressing behavior, children may develop the skills (including those related to conflict resolution) needed to more constructively approach challenging situations in life. Teachers (and schools) have been found to have meaningful, and varying, effects on behavioral outcomes, beyond test scores, for which we know there are meaningful returns (Jackson, 2018; Petek and Pope, 2021; Rose, Schellenberg and Shem-Tov, 2022). Restorative practices facilitate the accumulation of such non-traditional human capital; and we investigate the returns to this skill development and show they are meaningful. Indeed, social preferences may remain consistent and stable over time, such that habits formed early in life may influence the way people conduct themselves later in life (Chuang and Schechter, 2015). Restorative practices, while not a panacea, may provide a set of tools to help school administrators and teachers accomplish this goal.

References

Acosta, Joie, Matthew Chinman, Patricia Ebener, Patrick S. Malone, Andrea Phillips, and Asa Wilks. 2019. "Evaluation of a Whole-School Change Intervention: Findings from a Two-Year Cluster-Randomized Trial of the Restorative Practices Intervention." *Journal of Youth and Adolescence*, 48(5): 876–890.

Angel, Caroline M, Lawrence W Sherman, Heather Strang, Barak Ariel, Sarah Bennett, Nova Inkpen, Anne Keane, and Therese S Richmond. 2014. "Short-term effects of restorative justice conferences on post-traumatic stress symptoms among robbery and burglary victims: a randomized controlled trial." *Journal of Experimental Criminology*, 10(3): 291–307.

Augustine, Catherine, John Engberg, Geoffrey Grimm, Emma Lee, Elaine Wang, Karen Christianson, and Andrea Joseph. 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 Corporation.

Bacher-Hicks, Andrew, Stephen B Billings, and David J Deming. 2019. "The school to prison pipeline: Long-run impacts of school suspensions on adult crime." *National Bureau of Economic Research.*

Bowles, Samuel, and Herbert Gintis. 1976. *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life.* Basic Books.

Breiman, Leo. 2001. "Random Forests." *Machine Learning*, 45(1): 5–32.

Callaway, Brantly, and Pedro Sant'Anna. 2020. "Difference-in-Differences with multiple time periods." *Journal of Econometrics, forthcoming.*

Chicago Public Schools. 2020. "District Data | Chicago Public Schools."

Chuang, Yating, and Laura Schechter. 2015. "Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results." *Journal of development economics*, 117: 151–170.

de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 1688–1699.

Dreeben, Robert. 1967. "The contribution of schooling to the learning of norms." *Harvard Educational Review*, 37(2): 211–237.

Evertson, Carolyn M, and Carol S Weinstein. 2006. *Handbook of classroom management: Research, practice, and contemporary issues.* Routledge.

Fabelo, Tony, Michael D Thompson, Martha Plotkin, Dottie Carmichael, Miner P Marchbanks, and Eric A Booth. 2011. "Breaking schools' rules: A statewide study of how school discipline relates to students' success and juvenile justice involvement." *New York: Council of State Governments Justice Center.*

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

Fulkerson, Andrew. 2001. "The use of victim impact panels in domestic violence justice approach." *Contemporary Justice Review*, 4: 355–368.

González, Thalia. 2012. "Keeping Kids in Schools: Restorative Justice, Punitive Discipline, and the School to Prison Pipeline." *Journal of Law and Education*, 41(2): 56.

Gregory, Anne, Francis L. Huang, Yolanda Anyon, Eldridge Greer, and Barbara Downing. 2018. "An Examination of Restorative Interventions and Racial Equity in Out-of-School Suspensions." *School Psychology Review*, 47(2): 167–182.

Griffith, David, and Adam Tyner. 2019. "Discipline Reform through the Eyes of Teachers." *Thomas B. Fordham Institute*.

Hopkins, Belinda. 2003. *Just schools: A whole school approach to restorative justice.* Jessica Kingsley Publishers.

Jackson, C Kirabo. 2018. "What do test scores miss? The importance of teacher effects on non–test score outcomes." *Journal of Political Economy*, 126(5): 2072–2107.

Karp, David R, and Beau Breslin. 2001. "Restorative justice in school communities." *Youth & Society*, 33(2): 249–272.

Kauffman, James M, Patricia L. Pullen, Mark P. Mostert, and Stanley C. Trent. 2011. *Managing classroom behavior: A reflective case-based approach.* Pearson.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in neural information processing systems*, 30.

Kirk, David S, and Robert J Sampson. 2013. "Juvenile arrest and collateral educational damage in the transition to adulthood." *Sociology of education*, 86(1): 36–62.

Lai, Ijun. n.d. "Short-term impacts of Chicago's Suspensions and Expulsions Reduction Plan (SERP)." *Dissertation Paper*.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*, 604–620.

Losen, Daniel, Damon Hewitt, and Ivory Toldson. 2014. "Eliminating Excessive and Unfair Exclusionary Discipline in Schools Policy Recommendations for Reducing Disparities." *Policy Recommendations for Reducing Disparities*, 16.

Losen, Daniel J, Cheri L Hodson, Michael A Keith II, Katrina Morrison, and Shakti Belway. 2015. "Are we closing the school discipline gap?"

Losen, Daniel J, Paul Martinez, et al. 2020. "Lost opportunities: How disparate school discipline continues to drive differences in the opportunity to learn." *The Civil Rights Project at UCLA*.

McCold, Paul, and Benjamin Wachtel. 1998. "Community is not a place: A new look at community justice initiatives." *Contemporary Justice Review*, 1(1): 71–85.

McGarrell, Edmund F. 2001. *Restorative Justice Conferences as an Early Response to Young Offenders.* US Department of Justice, Office of Justice Programs, Office of Juvenile ....

Minow, Martha. 2022. "Restorative Justice and Anti-Racism." *Nevada Law Journal, Forthcoming*.

Parsons, Talcott. 1959. "The school class as a social system: Some of its functions in American society." *Harvard Educational Review*, 29: 297–318.

Petek, Nathan, and Nolan G Pope. 2021. "The multidimensional impact of teachers on students." Working Paper.

Pope, Nolan, and George Zuo. 2020. "Suspending suspensions: The education production consequences of school suspension policies."

Public Agenda Foundation, New York, NY. 2004. *Teaching Interrupted. Do Discipline Policies in Today's Public Schools Foster the Common Good?.* ERIC Clearinghouse.

Rose, Evan K, Jonathan T Schellenberg, and Yotam Shem-Tov. 2022. "The Effects of Teacher Quality on Adult Criminal Justice Contact." *National Bureau of Economic Research*.

Sartain, Lauren, Elaine M Allensworth, and Shanette Porter. 2015. "Suspending Chicago's Students." *The University of Chicago Consortium on Chicago School Research*, 74.

Shem-Tov, Yotam, Steven Raphael, and Alissa Skog. 2021. "Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program." *National Bureau of Economic Research*.

Shollenberger, Tracey L. 2015. "Racial disparities in school suspension and subsequent outcomes." *Closing the school discipline gap: Equitable remedies for excessive exclusion*, 31–44.

Stevens, W David, Lauren Sartain, Elaine M Allensworth, and Rachel Levenstein. 2015. "Discipline Practices in Chicago Schools." *The University of Chicago Consortium on Chicago School Research*, 52.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics, forthcoming*.

UChicago Impact. 2021. "5Essentials: Evidence-based school improvement."

Understood. 2023. "The Difference Between IEPs and 504 Plans."

Wadhwa, Anita. 2015. *Restorative justice in urban schools: Disrupting the school-to-prison pipeline.* Routledge.

Wang, Ke, Yongqiu Chen, Jizhi Zhang, and Barbara A Oudekerk. 2020. "Indicators of School Crime and Safety: 2019. NCES 2020-063/NCJ 254485." *National Center for Education Statistics*.

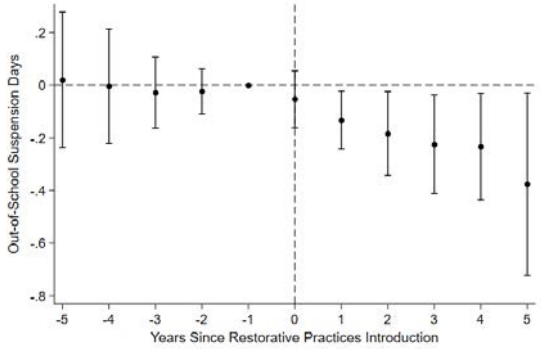
Winn, Maisha. 2016. "Transforming justice: Transforming teacher education." *Teaching Works*.

Wolf, Kerrin C., and Aaron Kupchik. 2017. "School Suspensions and Adverse Experiences in Adulthood." *Justice Quarterly*, 34(3): 407–430.

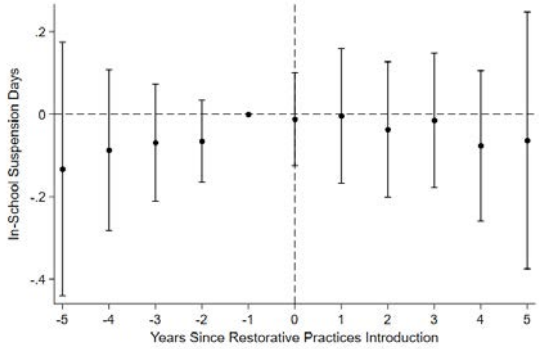
X Main Figures

Figure 1: High School Event Studies: Behavioral Outcomes

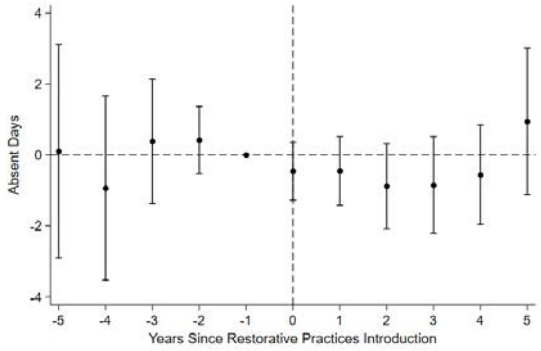
(a) Out-of-School Suspensions



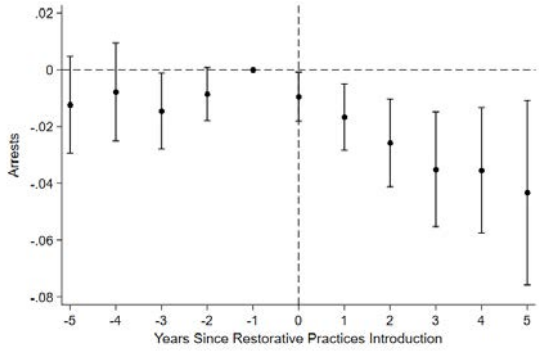
(b) In-School Suspensions



(c) Absent Days

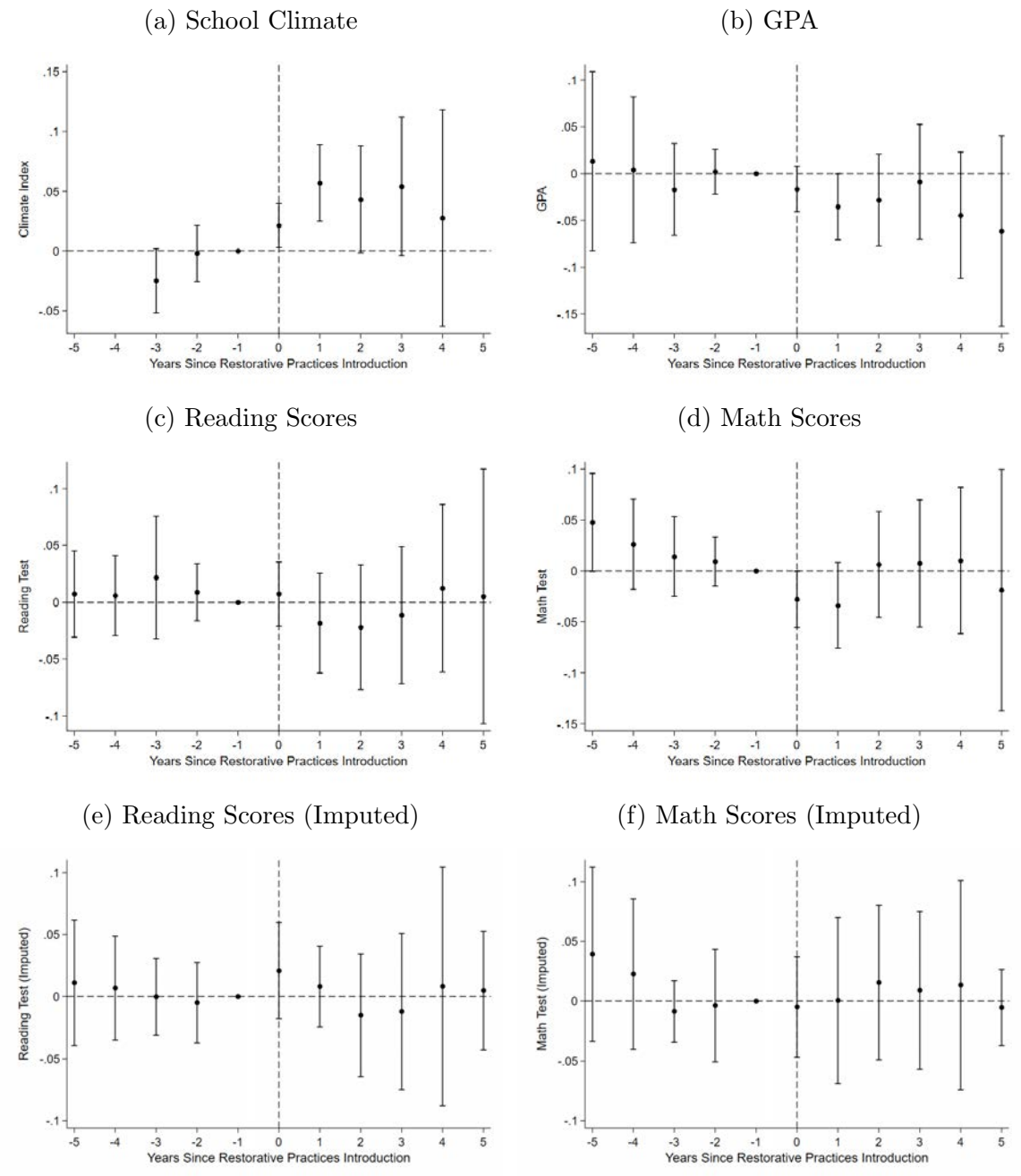


(d) Arrests



Notes: These figures show the event studies around the introduction of RP on in-school behavioral outcomes (out-of-school suspensions, in-school suspensions, and absent days) and policing outcomes (overall arrests) over time in high schools. Observations are at the student-school year level. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. Suspension and absence data are collected by Chicago Public Schools. An out-of-school suspension is defined as the removal of a student from class attendance or school attendance. An in-school suspension is defined as the removal of a student from their regular educational schedule for more than 60 minutes of the school day to an alternative supervised setting inside the school building. The absent days outcome is adjusted to equal total absent days minus out-of-school suspension days. Arrest data are collected by the Chicago Police Department. The arrest outcome is defined as the number of arrests experienced by students in a given year, regardless of the type of arrest or the location of the arrest. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Regressions for the absent days outcome include student member days in the corresponding school year as a control. Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Bars represent 95% confidence intervals based on standard errors clustered by school.

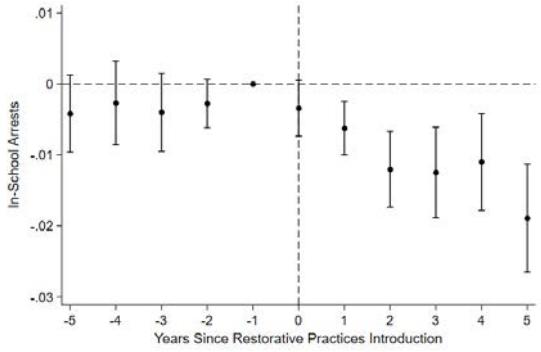
Figure 2: High School Event Studies: School Climate and Learning



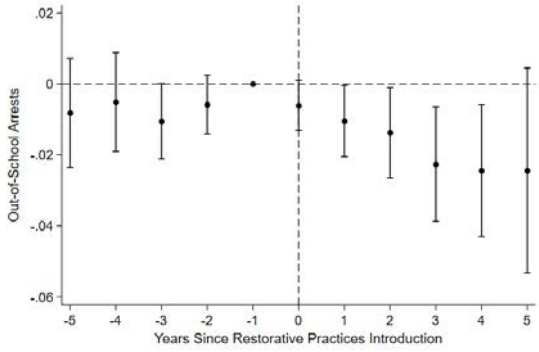
Notes: These figures show the event studies around the introduction of RP on students' perceptions of school climate and academic outcomes (GPA, reading test score, math test score, imputed reading score, and imputed math score) over time in high schools. Observations are at the student-school year level. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. The school climate index measures student socioemotional wellbeing levels and perceptions regarding the supportiveness of school environments based on constructs from the My Voice My School (MVMS) survey. Data for the school climate index begin two years after and ends one year before the data for the other outcome variables. Its graph therefore reflects one fewer estimated dynamic effect and two fewer placebo effects. GPA is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade; imputation is based on the methodology described in the text in Section VI. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Bars represent 95% confidence intervals based on standard errors clustered by school.

Figure 3: High School Event Studies: Policing Outcomes

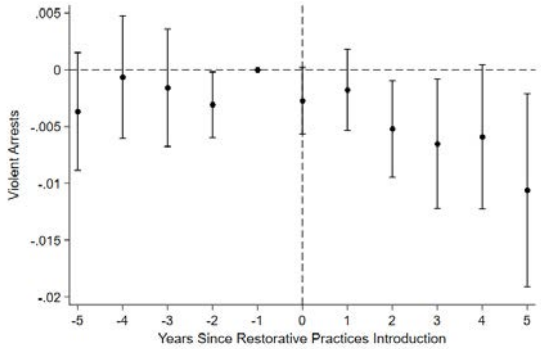
(a) Number of In-School Arrests



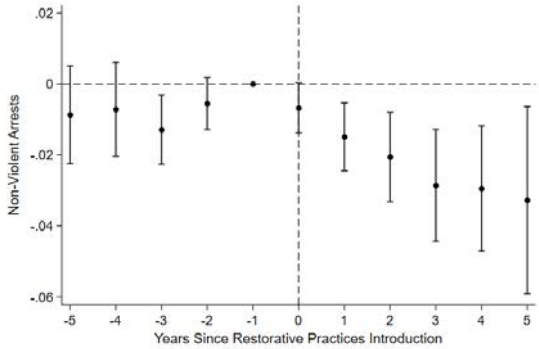
(b) Number of Out-of-School Arrests



(c) Number of Violent Arrests



(d) Number of Non-Violent Arrests



Notes: These figures show the event studies around the introduction of RP on students' arrest outcomes (out-of-school vs. in-school, and violent vs. non-violent) over time. Observations are at the student-school year level. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. Arrest data are collected by the Chicago Police Department. The arrest data includes information on the type (violent or non-violent), the location, and the time of arrest. In-school arrests are defined as incidents that happened both inside the school location and during school hours, and out-of-school arrests are defined as incidents that happened either outside the school location or outside school hours. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Bars represent 95% confidence intervals based on standard errors clustered by school.

XI Main Tables

Table 1: Baseline Characteristics: Chicago Public Schools High School Students

Variable	Treated (1)	Non-Treated (2)	Difference (3)
Number of Students	1003.69 (774.87)	448.92 (398.71)	554.78** (102.64)
Out-of-School Suspension Days	1.03 (3.20)	0.83 (2.80)	0.20 (0.18)
In-School Suspension Days	0.47 (1.67)	0.34 (1.53)	0.14 (0.12)
Absent Days	21.07 (20.88)	15.06 (17.97)	6.02** (1.52)
Number of Arrests	0.13 (0.62)	0.12 (0.63)	0.01 (0.03)
Ever Arrested	0.08 (0.27)	0.06 (0.24)	0.01 (0.01)
GPA	2.40 (0.98)	2.59 (0.98)	-0.19+ (0.11)
Math Scores	-0.09 (0.92)	0.12 (1.08)	-0.21 (0.15)
Reading Scores	-0.08 (0.94)	0.10 (1.06)	-0.18 (0.16)
School Climate Index	-0.07 (0.62)	0.10 (0.65)	-0.17** (0.05)
English Learner	0.07 (0.25)	0.05 (0.22)	0.02 (0.01)
Students in Temporary Living Situations	0.06 (0.24)	0.06 (0.24)	0.00 (0.01)
Individualized Education Plan	0.14 (0.34)	0.13 (0.34)	0.00 (0.01)
Economically Disadvantaged	0.84 (0.37)	0.81 (0.39)	0.02 (0.04)
Gender: Female	0.51 (0.50)	0.52 (0.50)	-0.01 (0.01)
Race: Black	0.41 (0.49)	0.50 (0.50)	-0.09 (0.08)
Race: White	0.10 (0.30)	0.08 (0.27)	0.02 (0.03)
Race: Latine	0.44 (0.50)	0.38 (0.49)	0.06 (0.06)
Disability: Cognitive	0.13 (0.33)	0.12 (0.33)	0.00 (0.01)
Disability: None	0.84 (0.37)	0.84 (0.37)	0.00 (0.01)
Disability: Physical	0.01 (0.10)	0.01 (0.10)	0.00 (0.00)
Disability: 504	0.03 (0.16)	0.03 (0.17)	0.00 (0.00)
Observations	58,784	44,214	

Notes: This table presents student-level means in subsequently treated high schools (column 1) and non-treated high schools (column 2), with means constructed in SY13 (prior to the introduction of RP). The associated differences (column 3) are derived from student-level regressions of the given outcome on a treatment indicator variable, with the standard errors clustered at the school level. Absent Days is defined as the total number of days absent, minus the total number of out-of-school suspension days that a student had in the school year, regardless of school. Arrest data are collected by the Chicago Police Department. GPA is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade. The School Climate Index measures student socioemotional wellbeing levels and perceptions regarding the supportiveness of school environments based on constructs from the My Voice My School (MVMS) survey. See Data Appendix C for detailed variable definitions. Standard errors are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 2: High School Restorative Practices: In-School Behavioral Outcomes

	Out-of-School Suspension		In-School Suspension		Absent Days
	Days	Binary	Days	Binary	
	(1)	(2)	(3)	(4)	(5)
RP	-0.167∗	-0.024∗	-0.028	-0.003	-0.540
	(0.068)	(0.010)	(0.068)	(0.019)	(0.484)
Baseline Mean	0.940	0.177	0.413	0.132	18.401
Observations	1,356,512	1,356,512	1,356,512	1,356,512	1,356,512

Notes: Observations are at the student-school year level, and we report the average effect of restorative practices over six periods. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. In columns 1 and 3, the out-of-school suspension (OSS) days and in-school suspension (ISS) days outcomes are the total number of OSS or ISS days that the student received in the corresponding school year, regardless of the school. In columns 2 and 4, the OSS and ISS binary outcomes indicate whether a student ever received either of these types of suspensions in the corresponding school year, regardless of the school. An out-of-school suspension is defined as the removal of a student from class attendance or school attendance. An in-school suspension is defined as the removal of a student from their regular educational schedule for more than 60 minutes of the school day to an alternative supervised setting inside the school building. In column 5, the absent days outcome is adjusted to equal total absent days minus out-of-school suspension days. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Regressions for the absent days outcome include student member days in the corresponding school year as a control. Data were collected by Chicago Public Schools. Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 3: High School Restorative Practices: Policing Outcomes

	Number of Arrests Overall	Number of In-School Arrests	Number of Out-of-School Arrests	Number of Violent Arrests	Number of Non-Violent Arrests
	(1)	(2)	(3)	(4)	(5)
RP	-0.024**	-0.009**	-0.015**	-0.005*	-0.020**
	(0.007)	(0.002)	(0.006)	(0.002)	(0.005)
Baseline Mean	0.128	0.026	0.102	0.028	0.100
Observations	1,380,959	1,380,959	1,380,959	1,380,959	1,380,959

Notes: Observations are at the student-school year level, and we report the average effect of restorative practices over six periods. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. Arrest data are collected by the Chicago Police Department. The arrest data includes information on the type (violent or non-violent), the location, and the time of arrest. The main arrest outcome is defined as the number of arrests experienced by students in a given year, regardless of the type of arrest or the location of the arrest. In-school arrests are defined as incidents that happened both inside the school location and during school hours, and out-of-school arrests are defined as incidents that happened either outside the school location or outside school hours. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 4: High School Restorative Practices: School Climate and Learning Outcomes

	School Climate (1)	GPA (2)	Reading Scores (3)	Math Scores (4)	Reading Scores (Imputed) (5)	Math Scores (Imputed) (6)
RP	0.042∗	-0.028	-0.005	-0.008	0.004	0.006
	(0.017)	(0.019)	(0.023)	(0.023)	(0.021)	(0.021)
Baseline Mean	0.000	2.457	0.000	0.000	-0.034	-0.036
Observations	751,792	897,230	831,928	824,298	942,925	942,465

Notes: Observations are at the student-school year level, and we report the average effect of restorative practices over six periods (five periods for the school climate index). Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. The school climate index measures student socioemotional wellbeing levels and perceptions regarding the supportiveness of school environments based on constructs from the My Voice My School (MVMS) survey. GPA is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade; imputation is based on the methodology described in the text in Section VI. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 5: High School Restorative Practices: Treatment Heterogeneity by Predicted Peer Group Suspension Days

	All Students			*Low Predicted OSS Days Students*		
	Out-of-School Suspension Days (1)	Reading Scores (2)	Math Scores (3)	Out-of-School Suspension Days (4)	Reading Scores (5)	Math Scores (6)
Below-Median Predicted OSS	-0.096	-0.031	-0.055	-0.005	-0.031	-0.056
	(0.060)	(0.037)	(0.043)	(0.042)	(0.038)	(0.043)
Observations	599,586	364,559	358,054	409,959	251,860	246,769
Above-Median Predicted OSS	-0.242∗	0.011	0.033+	-0.020	-0.007	0.031
	(0.108)	(0.023)	(0.019)	(0.044)	(0.031)	(0.024)
Observations	634,316	370,187	369,522	190,757	116,902	116,782
Control for Own Predicted Suspension				✓	✓	✓

Observations are at the student-school year level, and we report the average effect of restorative practices over six periods. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. See Data Appendix C for detailed variable definitions. We present results for students belonging to school-by-cohort cells that are above- versus below-median in predicted suspension days within a given cohort. Low predicted OSS days students are those with below-median predicted suspension days within a given cohort. Predictions for out-of-school suspension days for each student are constructed using a random forest algorithm as described in the text in Section VII. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 6: High School Restorative Practices: Race-by-Gender Treatment Heterogeneity

	Out-of-School Suspension Days (1)	Absent Days (2)	Number of Arrests (3)	School Climate (4)	Reading Scores (5)	Math Scores (6)
Black Female	-0.325**	-0.658	-0.015*	0.061*	0.013	0.020
	(0.110)	(0.597)	(0.007)	(0.029)	(0.023)	(0.019)
White Female	-0.075+	-0.869	-0.007+	0.007	-0.028	-0.038
	(0.039)	(0.802)	(0.004)	(0.031)	(0.046)	(0.041)
Latina Female	-0.053	-0.284	-0.004	0.048*	-0.034	-0.035
	(0.035)	(0.669)	(0.002)	(0.020)	(0.035)	(0.038)
Black Male	-0.384**	-1.655**	-0.079**	0.041	0.010	0.044*
	(0.118)	(0.572)	(0.021)	(0.027)	(0.023)	(0.022)
White Male	-0.042	-0.620	-0.008	0.041	0.027	0.006
	(0.072)	(0.829)	(0.016)	(0.029)	(0.052)	(0.050)
Latino Male	0.003	-0.051	-0.019*	0.017	-0.050+	-0.055
	(0.066)	(0.632)	(0.010)	(0.020)	(0.026)	(0.035)

Notes: This table shows results by student race and gender. Observations are at the student-school year level, and we report the average effect of restorative practices over six periods (five periods for the school climate index). Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.

Table 7: High School Restorative Practices: Treatment Heterogeneity by Race and School Share of Black Students

	Out-of-School Suspension Days	Reading Scores	Math Scores
Panel A: All Students			
Below-Median Black Student Share	-0.111+	-0.025	-0.055
	(0.066)	(0.032)	(0.037)
Observations	466,881	287,904	284,209
Above-Median Black Student Share	-0.381**	-0.021	0.022
	(0.104)	(0.026)	(0.022)
Observations	731,432	436,572	433,468
Panel B: Black Students			
Below-Median Black Student Share	-0.351**	-0.030	-0.047
	(0.117)	(0.043)	(0.045)
Observations	48,291	28,568	28,269
Above-Median Black Student Share	-0.387**	0.011	0.039+
	(0.126)	(0.022)	(0.020)
Observations	496,497	289,074	287,885
Panel C: Latine Students			
Below-Median Black Student Share	-0.049	-0.053+	-0.072+
	(0.059)	(0.031)	(0.041)
Observations	322,721	201,093	199,230
Above-Median Black Student Share	-0.038	-0.054	0.014
	(0.091)	(0.037)	(0.035)
Observations	170,047	107,526	106,363

Notes: A school is classified as above- or below-median Black student share based on SY13 enrollment and the median student's school-level share in SY13. The share of Black students for schools not observed in SY13 is calculated using the first year that the school appears in the sample. Observations are at the student-school year level, and we report the average effect of restorative practices over six periods. Student treatment assignment is determined by the first high school a student had been enrolled in since SY09, and the sample covers students in grades 9 to 12 between SY09 and SY19. See Data Appendix C for detailed variable definitions. Each specification includes the following covariates: student age fixed effects, student cohort fixed effects (based on grade and school year of entry), ELL indicator, unhoused indicator, IEP indicator, free or reduced-price lunch indicator, gender fixed effects, race fixed effects, and disability status indicators (having a 504 plan, physical disability, or cognitive disability). Estimates are based on the methodology developed in de Chaisemartin and D'Haultfoeuille (2020) and described in the text. Robust standard errors clustered by school are reported with ** denoting statistical significance at the 1 percent level, * at the 5 percent level, and + at the 10 percent level.