American Economic Review 2025, 115(8): 2722–2754
https://doi.org/10.1257/aer.20230971

From Retributive to Restorative: An Alternative Approach to Justice in Schools†

By Anjali Adukia, Benjamin Feigenberg, and Fatemeh Momeni*

School districts historically approached conflict resolution from the perspective that suspending disruptive students was necessary to protect their classmates, even if this caused harm to perceived offenders. Restorative practices (RP)—focusing on reparation, accountability, and shared ownership of disciplinary justice—are designed to address undesirable behavior without harming students. We study Chicago Public Schools' adoption of RP and find that suspensions and arrests decreased, driven by effects for Black students. We find null effects on test score value added, ruling out meaningful average declines. We estimate a 15 percent decrease in out-of-school arrests, consistent with RP substantively changing student behavior. (*JEL* D63, D74, D91, I21, I28, J15, J16)

Classroom management and discipline represent one of 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 in the hopes of minimizing disruption and with the goal of maintaining a safe and secure environment conducive for learning. In school year (SY) 2011–2012, approximately 3.5 million public school students were suspended from school in the United States, losing nearly 18 million days of instruction (Losen et al.

*Anjali Adukia, University of Chicago, NBER, IZA (email: adukia@uchicago.edu); Benjamin Feigenberg, University of Illinois at Chicago, NBER (email: bfeigenb@uic.edu); Fatemeh Momeni, University of Chicago Education Lab (email: fmomeni@uchicago.edu). Erzo F.P. Luttmer was the coeditor for this article. For helpful feedback, we thank Roseanna Ander, Hellen Antonopoulos, Monica Bhatt, Sandy Black, Jessika Bottiani, Jadine Chou, David Griffith, Jon Guryan, Rick Hess, Ariel Kalil, Jens Ludwig, Ben McKay, Martha Minow, Dick Murnane, Ben Ost, Rachel Perera, Mike Petrilli, Nolan Pope, Javaeria Qureshi, Steven Raphael, Stephen Raudenbush, Steve Rivkin, Jean Sack, Justina Schlund, Jeff Smith, Doug Staiger, Adam Tyner, Anita Wadhwa, Chezare Warren, staff members at the Chicago Public Schools, four anonymous referees, and seminar participants at AERA, AEFP, 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, The University of Chicago, Texas A and M University, and the Urban Institute. For excellent research support, we thank David Arbelaez, Bryant Cong, Jiahe Fan, 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 Arnold Ventures, 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. IRB approval was obtained from the University of Chicago (IRB18-0521). For data acquisition, we thank the Chicago Public Schools and the Chicago Police Department.
†Go to https://doi.org/10.1257/aer.20230971 to visit the article page for additional materials and author disclosure statement(s).

2015), with the rate of school suspensions more than doubling for Black and Latine children since 1974 (Losen and Martinez 2020).[1]

Many districts believe suspending disruptive students is essential to maintain order, promote accountability, and prevent negative spillover effects. In the canonical Becker (1968) framework, suspensions are intended to deter misbehavior by increasing punishment costs relative to more lenient responses. Being in a stricter school, however, can lead to long-term negative consequences such as decreased educational attainment, increased misconduct, and increased likelihood of incarceration (Fabelo et al. 2011; Shollenberger 2015; Wolf and Kupchik 2017; Bacher-Hicks, Billings, and Deming 2019). 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 support for the removal of students exhibiting misconduct from school premises to promote accountability (Public Agenda Foundation 2004). In recent years, a small but growing movement within education has sought solutions that constructively promote desirable behavior without relying on the threat of punitive discipline.

In our study, we investigate one such approach: 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; Riestenberg 2003; Mirsky 2007; Baker 2008; McCold 2008; Lewis 2009; González 2012; Angel et al. 2014; Anyon et al. 2014; Teasley 2014; González 2015; Wadhwa 2015; Winn 2016; Augustine et al. 2018; Gregory et al. 2018; Hashim, Strunk, and Dhaliwal 2018; Acosta et al. 2019; Shem-Tov, Raphael, and Skog 2024; Minow 2022).

We examine the impacts of RP by leveraging the rollout of RP programs across 73 high schools within the Chicago Public Schools (CPS) system beginning in SY14. Collectively, the 239 high schools in our study sample (including those that did not implement RP and that operated for only part of our study period) 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,

[1] For brevity, we will refer to school years by the year in which the spring term occurs (e.g., school year 2013–2014 is SY14), following CPS convention. We also refer to Black or African American children as Black children and Latine/a/o/x or Hispanic children as Latine children.

and juvenile arrests respond to RP exposure. In additional analyses, we examine outcomes at the elementary school level, where 214 out of 584 elementary schools introduced RP.

We find that RP decreased out-of-school suspensions by 17.8 percent for high school students. We do not find evidence of corresponding increases for 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, RP may be having a productive impact on teacher and/or student behavior. Teachers may change how they interact with students, better respond to students' individual needs, and avoid escalation. Students may resolve conflicts more effectively, understand their roles in conflicts, and 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 estimate a 18.8 percent overall decrease in child arrests, with declines for both violent (14.8 percent) and nonviolent (19.8 percent) offenses. These reductions in arrests reflect decreases during school hours and on school grounds (34.6 percent) as well as outside of school (14.7 percent). Declining in-school arrests may be driven by changes in how school staff respond to misconduct. In contrast, police officers serving outside of schools operate independently from school policies and practices, thus the decline in out-of-school arrests offers the strongest evidence of genuine changes in underlying student conduct.

Additionally, in accordance with the theory that RP may shift school culture, we find suggestive evidence of improved student perceptions of school climate based on student survey responses related to classroom behavior of peers, psychological sense of school membership, student-teacher trust, school-wide future orientation, and school safety.[2]

A common concern is that reduced punitiveness may lead to increased classroom disruption and resultant decreases in learning and academic achievement. There is mixed evidence on this question. On the one hand, Hinze-Pifer and Sartain (2018) and Craig and Martin (2023) find evidence indicating improved student outcomes following restrictions on exclusionary discipline in Chicago and New York alongside efforts to transform school cultures. By contrast, Pope and Zuo (2020) highlight the deficiencies of simply restricting teachers from using exclusionary discipline without providing alternative tools to address misconduct. They find suspension reduction policies in Los Angeles decreased suspension rates, but also led to declines in academic performance and increased absences and teacher turnover. In our setting, we do not find significant changes in learning outcomes following the introduction of RP. We can rule out math (reading) test score value-added declines larger than 0.013 (0.033) standard deviations based on 95 percent confidence intervals.

Evidence of improvements in students' perceptions of classroom behavior from student surveys also points against increases in classroom disruption. To more rigorously test for classroom disruption, we employ a random forest algorithm to classify

[2] We interpret estimated school climate impacts with caution given visual evidence that climate perceptions in RP-adopting schools may have begun to improve prior to adoption.

students based on their classmates' predicted suspension rates under the status quo disciplinary system, which we show in turn predicts suspension rate declines in response to the introduction of RP. Focusing on students who are themselves at low risk of suspension and therefore less likely to experience any suspension-related change in instructional time, we find no evidence of differential test score declines in schools with above-median predicted suspension rates. Although we lack the precision needed to confidently rule out meaningful differences in test score impacts as a function of peers' predicted suspension rates, our findings taken as a whole provide suggestive evidence that disruption effects are not of first-order concern in the study setting.

Finally, we investigate treatment effect heterogeneity with a focus on student race/ethnicity 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 RP. Black males in particular, who are suspended for four times as many days as White male students and arrested six times more frequently at baseline, experience the largest declines in out-of-school suspension days and arrests as well as significant attendance gains (above and beyond the increase associated with reduced suspension days).

Taken together, our findings suggest that RP has the potential to improve student perceptions of school climate and reduce behavioral incidents inside and outside of school without harming academic performance, potentially improving the daily experiences of all students, regardless of their predicted exposure to exclusionary discipline absent RP. Our work builds on recent experimental evidence that has highlighted the promise of employing RJ in the US juvenile justice system (Shem-Tov, Raphael, and Skog 2024; see Strang et al. 2013 for a summary of earlier work on RJ in the justice system context more broadly) as well as less clear-cut evidence from educational settings. Most closely related to our study is Augustine et al. (2018), which evaluates the effects of RP adoption based on a randomized trial in which 22 of 44 schools in Pittsburgh, Pennsylvania (concentrated at the elementary level) were randomly assigned to receive RP programming. The authors find that RP programming led to suspension reductions, with mixed findings related to school climate, no measured changes in arrests or violent offenses, and suggestive evidence of reductions in academic achievement.[3]

The rest of the paper proceeds as follows. In Section I, we describe a conceptual framework related to how RP may influence outcomes in schools. In Section II, we describe the policy setting. In Section III, we discuss the data we use to estimate impacts. In Section IV, we explain our research design and outline the value-added framework used to estimate impacts on test scores. In Section V, we discuss our

[3] CPS' RP trainings were notably more intensive than those offered in the context of Augustine et al. (2018), which may contribute to differences in findings across the two contexts. In another study, Acosta et al. (2019), 7 of 13 middle schools in Maine were randomly assigned to RP programming; the authors do not find any significant impacts on student perceptions of the school environment or their own self-reported experiences. By contrast, a large number of pre-post evaluations find promising associations between RP participation and a range of outcomes (see, for instance, McMorris et al. 2013). These studies contribute to a rich body of work on restorative justice that examines differences in approaches, settings, and outcomes (see McCold and Wachtel 1998; Fulkerson 2001; Karp and Breslin 2001; McGarrell 2001; Hopkins 2003; Riestenberg 2003; Mirsky 2007; Baker 2008; McCold 2008; Lewis 2009; González 2012; Angel et al. 2014; Teasley 2014; Winn 2016; Augustine et al. 2018; Gregory et al. 2018; Hashim, Strunk, and Dhaliwal 2018; Acosta et al. 2019; Shem-Tov, Raphael, and Skog 2024).

findings. In Section VI, we discuss possible disruption effects as a mechanism. In Section VII, we present treatment effect heterogeneity. In Section VIII, we conclude. Figures A1–A32 and Tables A1–A33 can be found in the Supplemental Appendix.

I. Conceptual Framework: Shaping Student Behavior in Schools

Consider a setting in which a student exhibits undesirable behavior ("the one who harmed," or the "offender") towards another individual ("the one who was harmed," or "victim") and school officials must decide how best to respond. In doing so, school officials aim to hold the offender accountable and ensure that they learn appropriate behavior for the future, while helping the victim to feel safe and to feel that justice has been served.

Exclusionary disciplinary responses, such as suspensions, temporarily remove the offender from the school setting and so may increase the victim's immediate feeling of safety and provide a reprieve from interacting with the offender. They also ideally give the offender time to reflect and provide a signal regarding appropriate behavior. However, isolation and deterrence alone may be insufficient to generate behavioral change (and some offenders may view time away from school positively). Moreover, victims often report that to feel justice has been served, they need offenders to take accountability for their actions and recognize the harm they caused. Simply removing the offender from school may fail to satisfy this objective. Exclusionary responses may also prove counterproductive to school officials' long-term goals by isolating students further from schools or increasing children's exposure to negative influences outside of school, which may perpetuate long-term harm through decreased educational attainment or increased criminal activity (Ross and Stillinger 1991; Fabelo et al. 2011; Shollenberger 2015; Wolf and Kupchik 2017; Bacher-Hicks, Billings, and Deming 2019). Such responses may also negatively affect parents' relationship with their children or their ability to work.

To achieve justice and accountability without generating the potential harms related to exclusion, school staff have turned to "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 preexisting relationship. In RJ, the different stakeholders are engaged through open dialogue with the goal of increased perspective taking, increased accountability, and shared ownership of disciplinary justice. The concept originated in indigenous practices and religious traditions. In modern times, school districts across the United States 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. RP can involve a conference between the offender and the victim, or between groups of victims or offenders who went through similar experiences. Each agent has to agree to whatever the process is; a victim will not be forced to participate if they feel doing so will retraumatize them or if they do not want to discuss their experiences. The precise structure of RJ is intentionally flexible and will vary based on the setting and situation.

Concretely, consider a situation of conflict mediation after one student assaults another student in response to a perceived slight. This incident may be addressed through restorative conversations with each student followed by a peace circle that involves the victim, the offender, and any bystanders. This process would allow the student(s) to explain the situation from their perspectives and to identify root causes of and harms caused by the incident as well as reflect on their immediate reactions, emotional response, and sense of what is needed for the harm to be repaired. The goal would then be to repair the harm done by determining logical consequences that are fair, sensible, and directly tied to the problematic behavior. For this example, such consequences could include attendance at conflict-resolution or anger-management workshops, a meaningful apology, role playing, or a written assignment that describes how the situation could have been handled more positively.[4] This emphasis on logical consequences that can serve to promote learning and self-reflection, as opposed to employing one-size-fits-all punitive disciplinary responses, 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 while maintaining a school environment that is conductive to learning. Whether RP achieves these objectives in practice is ultimately an empirical question.

II. Policy Setting: Chicago Public Schools

We study the impacts of RP in the context of CPS, one of the largest school districts in the United States, which serves over 340,000 students across more than 600 schools. The population of CPS is racially and economically diverse. Of the students attending CPS in SY25, 34.2 percent identified as Black, 47.3 percent as Latine, and 11.3 percent as White, and 71.6 percent were eligible to receive free or reduced-price lunches. Like many other large school districts in the 1980s and 1990s, CPS implemented 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 resultant high suspension rates, especially among students of color and among students from the most vulnerable backgrounds (Stevens et al. 2015; Sartain, Allensworth, and Porter 2015). In response, CPS explored alternative approaches designed to improve student safety and learning. This included the "Culture of Calm Initiative," which was launched in SY10. Specific program components widely varied across schools but included mentorship, job programs, socioemotional learning, and elements of RJ (Levenstein, Sporte, and Allensworth 2011; Zagar et al. 2013; Hinze-Pifer and Sartain 2018).

[4] If there are bystanders who actively or passively witnessed the incident, school officials may separately seek to make sure they feel safe and that they are deterred from exhibiting the undesirable behavior in the future. While suspensions may achieve these objectives, they are not designed to promote bystanders' agency and involvement in the event that future behavioral incidents arise. Indeed, prior research suggests that punitive approaches may foster a culture of abdication of responsibility or perpetuate victimization among bystanders (Twemlow, Fonagy, and Sacco 2004; Wilson-Simmons et al. 2006). An RJ-informed response might involve bystanders being asked to reflect on their roles in the incident and how they might help prevent such a situation from arising in the future.

In SY13, CPS implemented a number of changes to their student code of conduct, including removing the automatic 10-day suspension for certain student behaviors and adding recommendations for nonexclusionary disciplinary practices for all schools (Stevens et al. 2015). Then in SY14, CPS announced a disciplinary policy reform plan called the Suspensions and Expulsion Reduction Plan (SERP), aimed at decreasing the use of exclusionary discipline. This spurred various policy changes through the student code of conduct which included restrictions on the tiers of infractions that could result in suspensions as well as regulations related to suspension lengths and district administrator approval requirements. These efforts were specifically expected to reduce inequities in suspension rates by race/ethnicity and other student characteristics and are associated with improved student outcomes (Sartain, Allensworth, and Porter 2015; Hinze-Pifer and Sartain 2018; Lai 2018).

A. *Rollout of Restorative Practices Programs at CPS*

In SY14, CPS's Office of Social and Emotional Learning (OSEL) began to roll out district-wide RP programs. This initiative was meant to give teachers clear guidance on alternative tools to suspension and to improve the school environment. CPS received a grant from the US Department of Justice (DOJ) to introduce RP starting in 22 high schools and 34 elementary schools. By SY19, they expanded their RP programs to reach 279 schools, including 73 high schools and 206 distinct elementary schools (8 high schools that introduced RP also served elementary grade levels). CPS offered different RP programs, including RP Coaching, RP Leadership, and RP Peer Council.[5] Each program was based on fundamental RP principles: community building, social and emotional learning, accountability, healing and reparation of harm, and restorative systems and mindsets.

The most intensive, and most common, of these programs was RP Coaching. The prescribed model involved having a trained expert meeting with and coaching administrators and designated individuals from an existing "School Climate Team" to demonstrate and implement RP within their school. From the school-based School Climate Team, one to two RP Leads were chosen and made responsible for training other staff and serving as a champion for RP throughout the building. The other School Climate Team members who participated in RP trainings reflected the organizational composition of the school community and could include administrators, teachers, nonteaching staff, and family/student representatives. The RP coaches were initially drawn from 15 different providers with specialists who had expertise in restorative justice and how to adapt to different and dynamic school situations. Coaches came to schools and met with teachers, administrators, and other designated school staff a few times each week throughout the academic year. This flexible model was designed to allow for varied implementation, to serve as ongoing professional development, and to meet schools' evolving needs and abilities by developing a menu of restorative practices most appropriate for their specific context. Once the DOJ funding ended in SY16, CPS reduced the number of vendors from which they

[5] 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.

drew, reduced the frequency of in-school engagement to once weekly, and slowed RP rollout to new schools.

The other two RP programs were less common but had similar objectives to RP Coaching. The second program was RP Leadership, which entailed a lighter touch intervention in schools. While RP Leadership shared the same objectives as RP Coaching, trainings involved 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 RP in CPS high schools as a whole, although we also briefly examine heterogeneity by program intensity.

Schools were selected to receive RP 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" groups such as students with Individualized Education Programs (IEPs) and Black students, a school's climate indicators on the *My Voice, My School* (MVMS) survey—now known as the CPS *5Essentials* survey (Hart et al. 2021), school size, and input from those working directly with the schools (network specialists). Staff within CPS' OSEL were responsible for establishing these criteria used to prioritize schools for RP programming and for ultimately deciding which schools would be allocated programming in a given year. In conversations with the research team, OSEL staff emphasized that the stated criteria were intended to identify those schools which could benefit most from RP programming.

III. Data Sources and Sample

Our analyses draw on four primary sources of data: (i) RP programming information from CPS, (ii) CPS administrative data on students, (iii) CPS data on student responses to the MVMS survey, and (iv) Chicago Police Department (CPD) arrests data.

Restorative Practices Programming Data.—To determine the timing of treatment for students enrolled in a given school, we use programming data provided to us by CPS's OSEL. These data include records on which schools first received RP training in each academic year between SY14 and SY19 as well as the type of training received.

Student Administrative Data.—We use CPS's student-level administrative data from SY09 to SY19 for information on student-level outcomes and demographics (Chicago Public Schools 2009–2019). The outcome variables include records of in-school and out-of-school suspensions, attendance records, course grades (used to construct GPA), and reading and math test scores. The demographic information includes data on student race, gender, a proxy for economic disadvantage (eligibility for free or reduced-price lunch), unhoused status, 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 dataset includes information on student-level enrollment history. We use a unique student ID generated by CPS to link these records to school-level OSEL programming data files and construct a student-level measure of treatment exposure. We describe these data in more detail in Supplemental Appendix C.

School Climate Data.—Since SY11, CPS has annually administered the *My Voice, My School* (MVMS) survey to students in grades 6 to 12 to understand their experiences in the school environment.

The survey comprises 21 constructs, and we create a climate index using data from student responses to the following 8 constructs that may be directly affected by the introduction of RP: emotional health, student classroom behavior, academic personalism, psychological sense of school membership, personal safety, school-wide future orientation, school safety, and student-teacher trust (Hart et al. 2021).

Police Arrest Data.—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 a measure of particularly severe perceived misconduct (i.e., that resulted in arrest). These data include individual-level arrest records from July 1, 2008 through September 2, 2019 (Chicago Police Department 2008–2019). The arrest data include information on the type of offense (violent or nonviolent), the location, and the time of arrest. We investigate separately the impact of RP by arrest type and by arrest timing/location (which we use to classify arrests as "in-school" versus "out-of-school"). The CPD and CPS data files are joined using probabilistic matching over a child's name, date of birth, gender, and home address.

A. *Study Sample*

Our benchmark analyses include observations from students enrolled in any CPS traditional (district-run), contract, or charter high school between SY09 and SY19 for at least one day. We focus our main analyses 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, in SY13, 2.2 percent (5.7 percent) of high school students were arrested in (outside) CPS schools, compared to 0.4 percent (0.8 percent) of elementary school students in grades 3 to 8. Ex ante, the low baseline arrest rate in elementary schools is expected to limit 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.

Table 1 presents average characteristics for students enrolled in the 184 CPS high schools in our sample in operation in the school year prior to the roll-out of RP (SY13), separately for schools that did and did not receive RP programming at any point between SY14 and SY19.[6] This table shows high schools that received RP

[6] Tables A2 and A3 present average baseline (SY13) characteristics by demographic group and based on alternative sample partitions. 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 RP programming at any point between SY14 and SY19.

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

Variable	Treated (1)	Nontreated (2)	Difference (3)
Total enrollment	1,004 (775)	449 (399)	555 (103)
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	0.14 (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	2.41 (0.97)	2.63 (0.97)	−0.22 (0.11)
Math scores (std.)	−0.09 (0.92)	0.12 (1.08)	−0.21 (0.15)
Reading scores (std.)	−0.08 (0.94)	0.10 (1.06)	−0.18 (0.16)
Math value added	−0.06 (0.56)	0.08 (0.60)	−0.14 (0.04)
Reading value added	−0.04 (0.63)	0.05 (0.65)	−0.09 (0.04)
School climate index (std.)	−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/ethnicity: Black	0.41 (0.49)	0.50 (0.50)	−0.09 (0.08)
Race/ethnicity: White	0.10 (0.30)	0.08 (0.27)	0.02 (0.03)
Race/Ethnicity: 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 nontreated 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; value added is then constructed based on the methodology described in the text in Section IVA. 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. The School Climate Index is standardized by school year and grade. See Supplemental Appendix C for detailed variable definitions.

training differed from nontreated 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, more negative perceptions of their school climates, and lower test score value added (test score levels are also lower but estimated differences are imprecise).[7] Finally, treated high schools were more likely to use suspensions as disciplinary tools at baseline. Though differences are not statistically significant at conventional levels, students who enrolled in subsequently treated schools had on average 38.2 percent more in-school suspension days (0.47 versus 0.34) and 24.1 percent more out-of-school suspension days (1.03 versus 0.83) than those enrolled in never-treated schools. These important

[7] To ensure that our attendance measure is not mechanically correlated with our out-of-school suspension (OSS) days measure, we subtract OSS days from the total number of absences. In-school suspension is not considered an absence because the student is still in a supervised setting inside the school. The test score value-added measure is described in Section IVA.

differences in student attributes by future school treatment status are consistent with the prioritization of RP programming described above.

IV. Empirical Strategy

Since schools that received RP programming differ on various dimensions when compared to schools that did not, we employ a difference-in-differences-style research design that relies on a conditional exogeneity assumption requiring that expected *changes* over time in outcomes absent treatment are independent of RP programming assignment.

To estimate impacts from RP exposure, 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 determine student-level treatment exposure based on the first high school that each student was enrolled in within the CPS system, as well as the year and grade level in which that student enrolled in CPS. To guide thinking, if student i was enrolled in high school g from SY10 to SY12, and then moved to high school g' in SY13, the student's treatment exposure remains a function of the timing of RP rollout in school g. The 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.[8] We follow an analogous approach when analyzing outcomes for students in elementary schools.

Our identification assumption is that students enrolling in schools that did and did not adopt RP over a given period would have exhibited parallel trends in relevant outcomes in the absence of the rollout of the RP treatment. An extensive recent literature has highlighted that estimators derived from standard two-way fixed effects models in settings with 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 2021). In our setting, there are several reasons why this homogeneity requirement is unlikely to be satisfied. First, RP impacts may be a function of cumulative exposure if behavioral changes take time to manifest. Second, teachers' disciplinary practices, and school climate more generally, may evolve over time as the core principles of RP 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. Since standard two-way fixed effects models rely on already-treated groups when constructing counterfactuals, this anticipated treatment effect heterogeneity (which is ultimately borne out in the data) introduces bias if changes in outcomes in already-treated groups are themselves partly driven by the dynamic effects of the treatment. As shown in Sun and Abraham (2021), 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

[8] 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. Across analyses, we exclude the following observations: 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, and -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.

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 RP, 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 treatment effect heterogeneity is present. In our setting, this estimator uses only students first enrolled in not-yet-treated schools to predict counterfactual outcomes.

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 RP exposure of student i with assigned school g in school year t. We define an assigned school as the school in which a student first enrolls (regardless of whether they later transfer). Students are not themselves assigned to schools as CPS has a district-wide school choice system. RP programming was introduced across grade levels within adopting schools, and we classify each school as exposed to the RP treatment in all years after its introduction. While we cannot measure the degree to which schools continued implementing RP with fidelity in subsequent years, conversations with OSEL staff and RP coaches indicate that a substantial majority of schools did continue implementing RP throughout the study period.[9] 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 (the treatment value is one if the school had introduced RP, and otherwise equals zero). $Y_{g,t}$ is the average value of outcome Y in school year t for students assigned to school g. Then, the instantaneous effect of RP in year t is equal to the difference between (i) a weighted average of the school-specific changes in outcomes between school year $t-1$ and t in schools first treated in school year t and (ii) a weighted average of the school-specific changes in outcomes between $t-1$ and t in schools untreated through school year t. In the first (second) weighted average, the weight for school g is the share of all students in schools first treated in (untreated through) school year t assigned to school g. Formally,

$$(1)\quad DID_t = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}\left(Y_{g,t} - Y_{g,t-1}\right) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}\left(Y_{g,t} - Y_{g,t-1}\right).$$

As 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. The weight corresponding to each year t is

[9] Though highly imperfect due to substantial variability across schools in reporting practices, school-level data on the use of RP actions in response to misconduct incidents also suggests that fade-out of RP use was limited. Among schools with available data that adopted RP prior to SY19, the average school-level share of misconduct records including an RP action code was 20.4 percent in the year in which RP was adopted and was 30.5 percent by SY19; approximately one-third of RP-adopting schools decreased their reported use of RP actions. To the extent that a subset of schools transitioned away from RP, or RP was partially adopted in "untreated" schools due to the arrival of previously trained staff, our treatment effect estimates will represent lower bounds on the true causal impact of persistent RP exposure.

the share of all students observed in the year that their assigned school is first treated (N_S) who were observed in year t. Formally,

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

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. Since the parallel trends assumption must be evaluated for each outcome of interest, we present these placebo and dynamic estimates in event study plots for all outcomes subsequently analyzed in our main tables.

Across analyses, our benchmark models also include the following student-level covariates: age fixed effects, cohort fixed effects, gender fixed effects, race/ethnicity fixed effects, and indicators for unhoused status, English learner status, having an IEP or a 504 plan, cognitive or physical disability, and free or reduced-price lunch eligibility. We also control for yearly total "member days" (the sum of days a student is present in and absent from school) when absent days is the outcome of interest. To incorporate covariates, we regress differences in outcomes across periods (for schools untreated across these periods) on corresponding differences in group-level average covariate values and time fixed effects. We then residualize all observations based on coefficient estimates. The inclusion of covariates improves the precision of estimates in some instances, but does not alter the pattern of findings (we reproduce all main exhibits without controls in the Appendix). Given treatment is assigned at the school level, we cluster standard errors (which we construct via bootstrap) at the level of the school in which each student first enrolled. The event study plots for high school student outcomes, presented in Figures 1–3 and A1–A2, provide support for the parallel trends assumption for key outcomes (see Figures A3–A7 for event study plots without controls; plots for elementary school students are presented in Figures A8–A13). In subsequent analyses, we take the following weighted average of instantaneous and dynamic estimates to produce a single estimate of the causal effect of treatment on the treated for each outcome,

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

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). To avoid small cell sizes, k is set to 4 for the school climate outcome and to 5 for all other outcomes.

A. *Value-Added Approach*

Section IV characterizes our benchmark specification, in which each outcome is measured in levels. We next turn to a value-added approach to estimate test score impacts. This approach controls for lagged student test scores using a two-step procedure and offers two key benefits. First, given that lagged test scores are strong predictors of contemporaneous test scores, the value-added approach is expected to improve estimate precision.[10] Second, we find evidence of selection into test-taking in response to RP adoption, which may bias estimated impacts on test score levels given that students with missing scores have lower predicted scores based on observable characteristics. The value-added approach likely mitigates selection bias because the association between student-level lagged test score gains and contemporaneous test score missingness is weaker than the association between lagged test scores and contemporaneous test score missingness.[11] Though typical concerns related to endogenous sorting across classrooms are not relevant given our study design, the value-added approach does provide additional reassurance that cross-school sorting does not bias estimated impacts (see Section V for additional discussion).

In step one of the value-added approach, we construct test score residuals by estimating regression models of the following form:

$$A^*_{ist} = \alpha_{st} + \beta X_{ist} + \epsilon_{ist}. \tag{4}$$

Here, A^*_{ist} represents the normalized (math or reading) test score of student i in school s in school year t. α_{st} represent school-by-year fixed effects. X_{ist} includes the same student-level covariates included in our benchmark regression models. In addition, X_{ist} includes grade-level indicators, lagged cubic polynomials in math and reading scores, and interactions between lagged test score regressors and grade-level indicators.[12] After estimating the regression model, we construct a residualized measure (separately for math and reading), $\nu_{ist} = \hat{\epsilon}_{ist} + \hat{\alpha}_{st}$. This measure, which captures the contribution of school s in school year t to test scores as well as the idiosyncratic component of student i's test score performance, serves as the dependent variable in the models we estimate in step two (where we use the same difference-in-differences-style estimator as described in Section IV).

V. Main Results

We seek to understand how school behavioral policies may shape 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.

[10] Lagged values are less predictive for other key outcomes, such as suspension days and arrests, given the sparseness of these measures.

[11] The effect of RP adoption on value-added missingness (which requires only a missing contemporaneous *or* lagged score) is also more muted than the effect of RP adoption on contemporaneous test score missingness (though estimates are not statistically distinguishable).

[12] The inclusion of school-by-year fixed effects mirrors the inclusion of teacher fixed effects in Chetty, Friedman, and Rockoff (2014) when the authors estimate β coefficients on student and classmate characteristics.

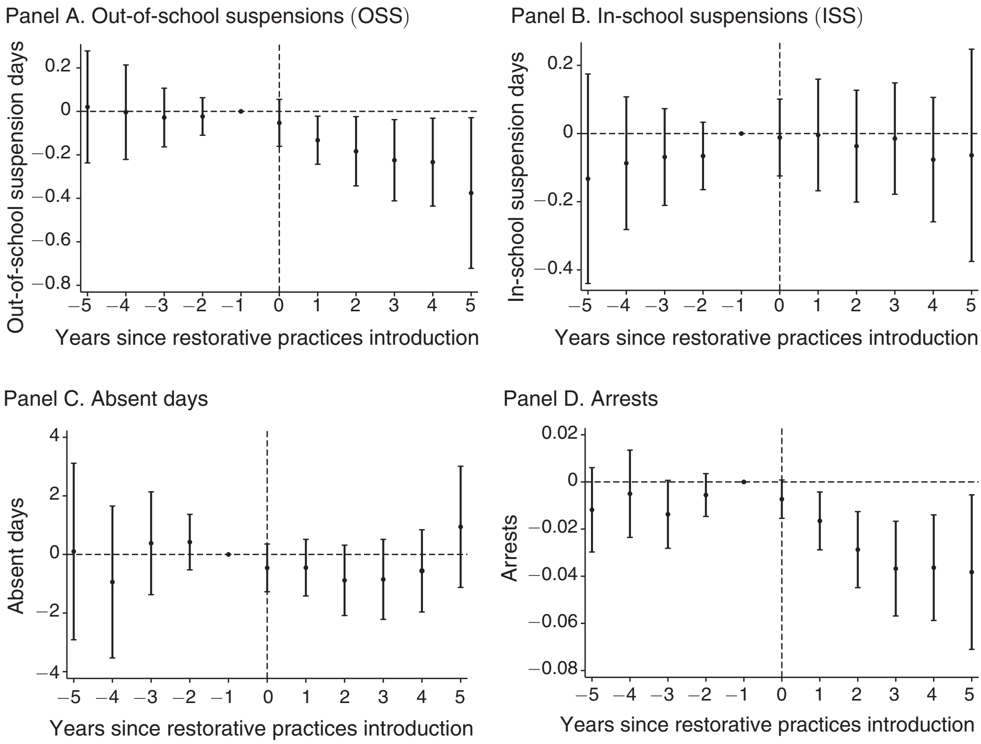

Figure 1. High School Event Studies: Behavioral Outcomes

Notes: These panels 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 Supplemental 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/ethnicity 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) described in the text. Bars represent 95 percent confidence intervals based on standard errors clustered by school and described in the text.

Changing School-Based Behavioral Outcomes.—First, we examine the impact of the introduction of restorative practices on suspensions and attendance. 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 estimate a significant decrease in out-of-school suspensions (OSS) of 0.17 days, or 17.8 percent (Table 2, column 1). This serves as evidence of a "first stage:" RP changed the behavior of teachers and/or students. By contrast, we estimate a relatively precise null impact on days absent and a noisier

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

	Out-of-school suspension		In-school suspension		
	Days	Binary	Days	Binary	Absent days
	(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 Supplemental 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/ethnicity 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. Standard errors are clustered at the school level.

null effect on in-school suspension days (Figure 1 and Table 2, columns 3, 4, and 5). Although only the total number of in-school suspension days is recorded in administrative records, the nature or duration (full- or part-day) of in-school suspensions may also have changed in response to RP exposure, with OSEL staff noting that RP training encouraged more productive uses of in-school suspension time (for instance, encouraging deescalation practices rather than having students pass the time in silence). In any case, these findings suggest that students are receiving more in-school instruction time, on average, in response to RP adoption.

Changing Behavior beyond the School.—We are interested in understanding whether being exposed to RP affects conflict resolution regardless of location and separate from structured or guided intervention. To do so, we draw on arrest data from the Chicago Police Department (CPD). Given prior evidence that student arrests are associated with worse long-term outcomes (Kirk and Sampson 2013), understanding the nature of changes in juvenile arrests helps elucidate the full extent of RP impacts.

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 RP. The estimated aggregate impact is an average decrease of 0.024 arrests, which represents a 18.8 percent decline relative to the baseline mean (Figure 1 and Table 3, column 1 of panel A). In column 1

Table 3—High School Restorative Practices: Policing Outcomes

	Arrests (overall) (1)	In-school arrests (2)	Out-of-school arrests (3)	Violent arrests (4)	Nonviolent arrests (5)
Panel A. Arrest outcomes (counts)					
RP	−0.024	−0.009	−0.015	−0.004	−0.020
	(0.007)	(0.002)	(0.006)	(0.002)	(0.005)
Baseline mean	0.128	0.026	0.102	0.027	0.101
Observations	1,380,959	1,380,959	1,380,959	1,380,959	1,380,959
Panel B. Binary arrest outcomes					
RP	−0.009	−0.006	−0.006	−0.003	−0.009
	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
Baseline mean	0.071	0.022	0.057	0.023	0.058
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 nonviolent), the location, and the time of arrest. The main arrest outcome is defined as the number of arrests (in panel A) or an indicator for any arrest (in panel B) 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 (outside of 7:00AM-6:59PM on school days). See Supplemental 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/ethnicity 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. Standard errors are clustered at the school level.

of panel B, we replace the arrest count dependent variable with an indicator for any arrest. We estimate a 12.7 percent decline in the likelihood of any arrest, relative to the baseline mean. This pair of estimates is consistent with relatively uniform percentage wise decreases in arrest counts across the baseline arrest count distribution (over half of those with any arrest at baseline were arrested exactly once).[13]

While the estimated decline in arrests in response to the introduction of RP is consistent with improved student behavior, school staff are tasked with referring students to law enforcement when 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 that adopt RP are encouraged to employ alternatives to traditional punitive approaches (including requests for law enforcement involvement) when possible. To distinguish between alternative explanations for the aggregate decline in student arrests, we next examine impacts on in-school arrests (defined initially based on whether the arrest takes place at the school location and between 7:00AM and 6:59PM on school days) versus out-of-school arrests (all other arrests). Since police officers serving outside of schools are not under the same authority as teachers and operate independently from school policies and

[13] The remaining columns of panel B of Table 3 present estimates for the binary versions of the arrest measures introduced below. Across outcomes, we find a similar ratio of count-based to binary outcome-based percentage wise effects.

practices, changes in out-of-school arrests can better capture genuine changes in student behaviors and approaches to conflict resolution.

In Table 3, columns 2 and 3 of panel A, we provide evidence that aggregate arrest declines reflect decreases for 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 RP. However, if out-of-school arrests occur disproportionately during school hours and outside of the school location on days when students are absent or suspended, it remains possible that we could find a decline in out-of-school arrests even in the absence of any behavioral change (through an incapacitation-type channel). To probe this possibility, in Table A5 we separately examine arrests outside of school hours. We find that such arrests decline by 14.0 percent, providing further support for hypothesized changes in student behavior. Although identifying the exact mechanism is beyond the scope of this study, these findings align with the hypothesis that RP has equipped students with conflict resolution skills that they can now apply beyond the school setting.[14]

A broader question is whether a restorative justice approach to conflict can decrease violence. To explore this question, we examined changes in arrests separately for violent and nonviolent offenses. We see declines in arrests for both types of offenses: a 14.8 percent reduction in the number of arrests for violent offenses and a 19.8 percent reduction in the number of arrests for nonviolent offenses (Table 3, columns 4 and 5 of panel A), suggesting that the introduction of RP also led to a decrease in violence.

Changing School Climate.—We saw that the introduction of RP resulted in a decrease in out-of-school suspensions (Table 2, columns 1 and 2). 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 this hypothesis, we find suggestive evidence of improvements in student-reported measures of school climate (Table 4). Specifically, we estimate a significant 0.042 standard deviation improvement in perceived school climate, though the negative placebo estimate from the third year before RP adoption does raise some concern regarding the validity of the parallel trends assumption for this outcome (see Figure 2). The climate index impact we estimate is driven by particularly large improvements in students' perceptions of their peers' classroom behavior, their psychological sense of school membership, their school-wide future orientation, and school safety (Table A6; Figure A15). 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

[14] Given that only 0.2 percent of all arrests take place in school between 5PM and 6:59PM, our preferred measure defines the school day as 7AM–4:59PM; we present results that alternatively use this cutoff and a more conservative 6:59PM cutoff. Corresponding event studies are presented in Figure A14.

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

	School climate (std.)	GPA	Reading value added (std.)	Math value added (std.)
	(1)	(2)	(3)	(4)
RP	0.042	−0.024	−0.002	0.016
	(0.017)	(0.021)	(0.016)	(0.015)
Baseline mean	0.000	2.473	0.000	0.000
Observations	751,792	851,492	421,783	421,864

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. The school climate index is standardized by school year and grade. GPA is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade; value added is then constructed based on the methodology described in the text in Section IVA. See Supplemental 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/ethnicity 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. Standard errors are clustered at the school level.

RP. These findings suggest that students are not only more likely to attend school, but that they also have more positive experiences while there.

Examining Student Learning.—Turning to academic outcomes, the estimated impact of RP adoption on student GPA is negative but not statistically significant at conventional levels. We estimate a 0.024 point decline in GPA and can only reject GPA declines larger than 0.07 points, or roughly 0.07 SD (Table 4, column 2). Given the school-level nature of the treatment we analyze, it is worth emphasizing that any nonzero impact of RP adoption on GPA would require a shift in the entire school-level GPA distribution.[15]

We next use the value-added framework described in Section IVA to analyze the impacts of RP adoption on student test scores. Despite evidence of improvements in student behavior and in school climate perceptions, we do not see any corresponding evidence of increased reading or math test score growth (Table 4, columns 3 and 4). Estimated impacts on test score gains in reading and math are small in magnitude and inconsistent in sign (−0.002 SD and 0.016 SD, respectively). Based on 95 percent confidence intervals, we can rule out reading (math) value-added declines that are larger in magnitude than 0.033 SD (0.014 SD).

[15] Note that this analysis excludes students enrolled in charter schools, which are not required to submit GPA records to CPS' Central Office. For reference, Kraft (2020) analyzes a sample of 747 randomized controlled trials that evaluate education interventions with standardized test outcomes. The median effect size is 0.10 SD in this sample, and the author classifies effects smaller than 0.05 SD as small.

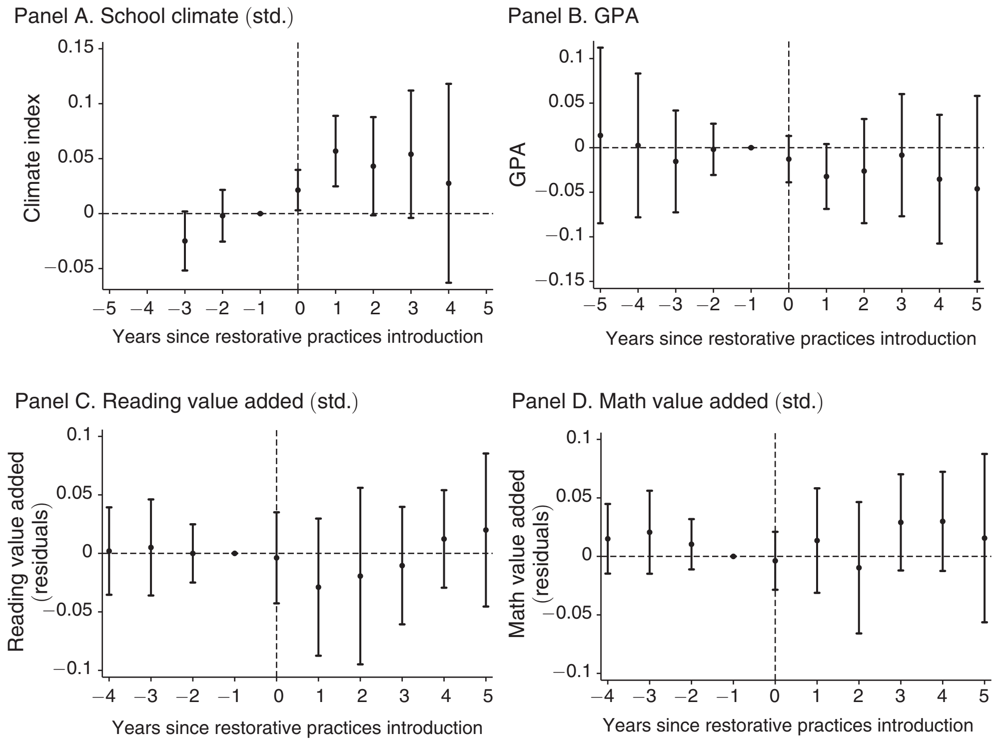

FIGURE 2. HIGH SCHOOL EVENT STUDIES: SCHOOL CLIMATE AND LEARNING

Notes: These figures show the event studies around the introduction of RP on students' perceptions of school climate and academic outcomes (GPA, reading value added, and math value added) 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. The school climate index is standardized by school year and grade. 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 (grade point average) is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade; value added is then constructed based on the methodology described in the text in Section IVA. Value added cannot be constructed for SY09; value-added graphs therefore reflect one fewer estimated placebo effect. See Supplemental 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/ethnicity 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 percent confidence intervals based on standard errors clustered by school.

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 find 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 A6). Nonetheless, in Section VI, we directly test for the

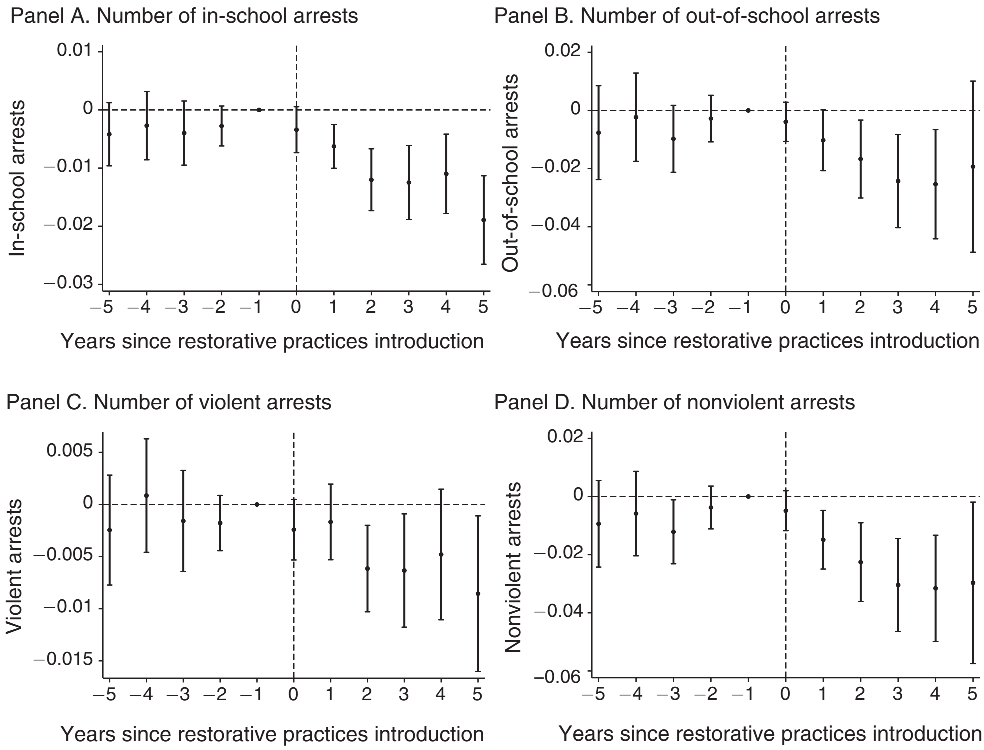

Figure 3. High School Event Studies: Policing Outcomes

Notes: These figures show the event studies around the introduction of RP on students' arrest outcomes (out-of-school versus in-school, and violent versus nonviolent) 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 nonviolent), 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 (outside of 7AM–6:59PM on school days). See Supplemental 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/ethnicity 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 percent confidence intervals based on standard errors clustered by school.

presence of disruption effects. We then investigate treatment effect heterogeneity by student characteristics to further unpack our average findings.

Additional Sensitivity Analyses.—We investigate the sensitivity of results to a range of alternative empirical approaches and specifications. We confirm that results remain robust across these alternative modeling 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 large estimated decline in absent days and corresponding evidence of differential pre-trends for this outcome (Tables A7–A8, panel B; Figures A16–A17).

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. Since charter and contract schools have some autonomy to establish their own Student Codes of Conduct and are not bound by the same administrative reporting obligations as district-run schools, we check the sensitivity of our results to excluding all observations for students who ever attended a charter or contract school in a given school year. The results remain largely unchanged (Tables A7–A8, panel C).

Excluding Controls: We verify that results are not sensitive to the exclusion of covariates by reproducing all main tables and figures with covariates excluded. We find qualitatively similar results (Tables A9 through A14; Figures A3 through A7).

Synthetic Difference-in-Differences: To probe the sensitivity of findings to the approach used to construct counterfactual outcomes, Tables A15–A16 and Figures A18–A19 present results derived using a synthetic difference-in-differences estimator (Arkhangelsky et al. 2021). We arrive at qualitatively similar conclusions with larger estimated impacts for key outcomes being driven primarily by differences in weighting (estimates based on de Chaisemartin and D'Haultfoeuille 2020, also increase in magnitude when all schools receive equal weight).

Additional Threats to Validity.—We investigate whether changes in enrollment or attrition patterns threaten the interpretation of findings.

Enrollment: Figure A20 demonstrates that schools that adopted RP were experiencing relative declines in enrollment prior to adoption and continued to experience differential enrollment declines in the post-adoption period. While the event studies we present for each key outcome provide direct support for our parallel trends-style identification assumption, here we present supplementary empirical tests to buttress the causal interpretation of findings. First, we examine whether the characteristics of students enrolling in RP-adopting schools vary with event timing (Table A17; Figure A21). We find little evidence that student demographics or predicted out-of-school suspension days are changing as a function of event time. Second, we reestimate models for all of our benchmark (non-test score) outcomes that are measured in levels while controlling for lagged values (see Table A18). If falling enrollment leads to more positive selection in schools that adopt RP, then controlling for lagged outcomes may substantially attenuate estimated RP impacts. In practice, this does not appear to be the case. Third, we estimate RP impacts with school-by-cohort (as opposed to school) as the grouping variable so that student composition is held fixed in the absence of CPS exit (we include only students who enroll in CPS by grade 9). Since cohorts enrolled entirely after RP adoption no longer contribute to treatment effect estimation, we expect estimates may both attenuate and lose precision due to sample size reductions and in the presence of treatment effects that grow over time. In practice, however, we arrive at qualitatively similar conclusions (see

Table A19). While the estimated impact on out-of-school suspension days decreases by nearly 60 percent in magnitude, estimated impacts on arrests and school climate perceptions closely mirror benchmark estimates in terms of magnitude (and precision increases marginally). The fact that impacts on arrests and school climate perceptions are unchanged also provides some reassurance that associated impacts are not mechanically related to benchmark suspension rate declines.

Attrition: We next test explicitly for differential attrition in order to understand the potential for selection bias on this margin (recall that a student who is not enrolled in any CPS school in a given year is absent from our study sample). 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 an attrition indicator as the dependent variable, we find little evidence of differential attrition. As shown in Figure A22, we estimate a small decline in attrition that is not statistically distinguishable from zero; we can rule out differential declines in attrition greater than 1.8 percentage points based on 95 percent confidence intervals.[16]

Elementary School Results.—RP was also introduced across elementary schools in CPS. In our main analyses, we focus on high school student outcomes because of low arrest rates and limited school climate data at the elementary school level, but it is still interesting to understand how RP influences outcomes among younger children. In Table 5, we present impacts for RP exposure among elementary school students. In column 1, we estimate a significant (12.5 percent) decline in out-of-school suspension days though this estimate should be interpreted cautiously given the positive placebo estimates presented in Figure A8. We estimate a null effect on in-school suspension days in column 2, with a 95 percent confidence interval that includes an increase of up to 0.017 days (a 31.5 percent increase given the rarity of in-school suspensions at the elementary level). The null estimate on absent days in column 3 allows us to rule out an increase in absent days greater than 0.11 days (1.3 percent). Turning to academic outcomes in columns 4–6, point estimates are inconsistent in sign and 95 percent confidence intervals allow us to rule out declines in GPA greater than 0.020 points (out of 4) and declines in reading (math) test score value added greater than 0.007 (0.008) SDs. Despite the low incidence of arrest among elementary school students, in column 7, we estimate a significant (18.3 percent) decline in arrests that closely mirrors our estimate for the high school sample (in percentage terms). Partitioning arrests based on location and timing, we estimate a significant 20.8 percent decline in out-of-school arrests along with an insignificant 16.0 percent decline in in-school arrests (Tables A20–A21; Figure A23).

[16]Examining behavioral outcomes, we find little evidence of differential selection into attrition by school RP adoption. Panels C–D of Figure A22 show that attritors in schools that do and do not implement RP are more likely by a similar magnitude to be suspended and arrested in their first year in a CPS high school than their nonattriting peers (the availability of first year data is not affected by subsequent attrition).

TABLE 5—ELEMENTARY SCHOOL RESTORATIVE PRACTICES: IN-SCHOOL BEHAVIORAL, LEARNING, AND POLICING OUTCOMES

	Out-of-school suspension days (1)	In-school suspension days (2)	Absent days (3)	GPA (4)	Reading value added (std) (5)	Math value added (std) (6)	Arrests (count) (7)	Arrests (binary) (8)
RP	−0.050	0.007	−0.077	−0.001	0.007	0.008	−0.0033	−0.0022
	(0.019)	(0.005)	(0.095)	(0.010)	(0.007)	(0.008)	(0.0014)	(0.0006)
Baseline mean	0.401	0.054	8.497	2.970	0.000	0.000	0.018	0.011
Observations	2,536,517	2,536,517	2,536,517	2,128,882	1,807,421	1,808,004	2,546,569	2,546,569

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 elementary school a student had been enrolled in since SY09, and the sample covers students in grades 3 to 8 between SY09 and SY19. In columns 1 and 2, 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. Suspension 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. In column 3, the absent days outcome is adjusted to equal total absent days minus out-of-school suspension days. GPA is calculated using semester final grades. Math and reading scores are standardized by test, school year, and grade; value added is then constructed based on the methodology described in the text in Section IVA. The column 7 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. The column 8 arrest outcome is an indicator for any arrest experienced by students in a given year, regardless of the type of arrest or the location of the arrest. Arrest data are collected by the Chicago Police Department. See Supplemental 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/ethnicity 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. Standard errors are clustered at the school level.

VI. Mechanisms: Disruption Effects

A key concern among those who advocate for more punitive disciplinary practices is that those students who are suspended under the status-quo system but who are less likely to be suspended after RP adoption will disrupt the learning of their peers. Then, null average impacts on academic outcomes could mask offsetting effects whereby students at risk of suspension benefit directly through increased engagement and an increase in instructional time, while those who were suspended at low rates at baseline (and so mechanically stand to benefit less on this margin) may be harmed academically. To test this hypothesis, we exploit variation in student-level exposure to potentially disruptive peers.

Employing a random forest algorithm, we first use data from SY13 and earlier to predict high school out-of-school suspension (OSS) days based on eighth-grade student characteristics (race/ethnicity, gender, number of arrests, attendance, GPA, and OSS days) as well as characteristics measured contemporaneously in high school (free or reduced-price lunch eligibility, English learner status, and unhoused status).[17]

[17] Relying only on pre-period data 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. For observations corresponding to SY14 and later, we use the random forest algorithm results (based on pre-period data) and student characteristics to predict OSS days.

Table 6—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 value added (std) (2)	Math value added (std) (3)	Out-of-school suspension days (4)	Reading value added (std) (5)	Math value added (std) (6)
School × Cohort Predicted to Have Below-Median OSS	−0.096	−0.029	−0.007	−0.059	−0.034	−0.020
	(0.066)	(0.027)	(0.025)	(0.053)	(0.026)	(0.026)
Observations	567,772	182,971	182,920	476,808	155,331	155,281
School × Cohort Predicted to Have Above-Median OSS	−0.290	0.018	0.035	0.020	-0.045	0.011
	(0.103)	(0.018)	(0.021)	(0.078)	(0.035)	(0.042)
Observations	658,618	192,518	192,532	168,621	51,722	51,725
Test (Above-Median = Below-Median): *p*-value	0.090	0.166	0.246	0.378	0.814	0.576
Control for own predicted suspension days				✓	✓	✓

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. See Supplemental 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. Students with low predicted OSS days 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 VI. 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/ethnicity 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. Standard errors are clustered at the school level.

By predicting baseline OSS days (as opposed to the effect of RP on OSS days), we rely on the testable hypothesis that RP-induced suspension declines will be largest where predicted baseline OSS days are highest.[18]

To classify students (both prior and subsequent to SY13) based on classmates' predicted baseline OSS days, we construct predicted OSS-day averages at the school-by-cohort level. We then split school-by-cohort cells into above- and below-median groups within a given cohort. We refer to these groups as "above-median" and "below-median" for brevity. Finally, we reestimate our benchmark regression models separately for students in above- versus below-median predicted suspension-day cells. The results, presented in Table 6, column 1, validate the use of predicted OSS days to generate heterogeneity in RP-induced OSS day declines. Students in above-median cells experienced a 0.290 day decline in OSS

[18] The random forest approach allows for arbitrary interactions between included covariates and relaxes 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 dataset. 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). See Breiman (2001) for further details on the bagging involved in the 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 features for each tree to split on equals the number of predictors divided by 3.

days in response to adoption (73.7 percent larger than our full sample estimate) compared to students in below-median cohort-by-school cells, who experienced a 0.096 day decline in OSS days (this difference is significant at the 10 percent level). In columns 2 and 3, we present estimated test score impacts for students in above- and below-median cells. Point estimates are negative for below-median students and positive for above-median students, though estimated impacts are not statistically distinguishable across the two subgroups.

To test for disruption effects, we next limit the sample to students with below-median predicted OSS days (with median values again constructed within cohort), who are unlikely to be suspended under either disciplinary regime. Indeed, we show in Table 6, column 4 that these students experience small and statistically insignificant changes in OSS days in response to RP adoption (point estimates for students in below- and above-median predicted suspension day cells are −0.059 and 0.020, respectively).[19] We then 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 6, we do not find evidence of heterogeneous test score impacts consistent with disruption effects. For students with below-median predicted classmate OSS days, we estimate an insignificant 0.034 SD decline in reading test score value added and an insignificant 0.020 SD decline in math test score value added. For students with above-median predicted classmate OSS days, we find an insignificant 0.045 SD decrease in reading test score value added and an insignificant 0.011 SD increase in math test score value added. p-values on the test of equality of reading and math estimates across subgroups are 0.814 and 0.576, respectively.[20] Although we cannot reject meaningful differences in test score impacts across subgroups (95 percent confidence intervals exclude only differential test score declines greater than 0.08 SD and 0.11 SD in math and reading), the pattern of findings also does not provide evidence in support of the disruption hypothesis.[21]

VII. Treatment Effect Heterogeneity

To understand the distributional implications of the average impacts we estimate, we consider treatment effect heterogeneity with an emphasis on differential impacts by student race/ethnicity and gender. For each source of heterogeneity analyzed, we conduct subsample-specific analyses and contrast treatment effect estimates (i.e., to

[19] An alternative approach would be to compare all students in below- versus above-median predicted OSS day cells while conditioning on their own predicted OSS days. In practice, however, we find that students in above-median predicted OSS day cells experience larger declines in OSS days in response to RP adoption, conditional on own predicted OSS 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.

[20] Table A22 presents baseline outcomes for students in above-median versus below-median cells. Figures A24 and A25 present event studies for out-of-school suspension days and test score outcomes by classmates' predicted suspension rates.

[21] An alternative (less direct) approach to estimating disruption effects is to assume that estimated test score impacts for low predicted OSS-day students provide an upper bound on the magnitude of disruption effects (since test score impacts should in theory capture disruption as well as any loss of learning due to time dedicated to RP). Under this assumption, we can rule out disruption-induced losses in test-score value added greater than 0.05 SD in math and 0.07 SD in reading based on 95 percent confidence intervals from regressions that estimate RP impacts in the pooled sample of students with low predicted OSS days.

investigate heterogeneity by student race and gender, we separately estimate benchmark regression models using the subsample of Black males, Black females, etc.).

Heterogeneity by Race/Ethnicity and Gender.—Student race/ethnicity 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 suspension (OSS) days in response to RP adoption, and we find that the aggregate reductions in OSS days we estimate are driven by Black male and female students, who experience declines of 0.384 and 0.325 suspension days, respectively (these estimated impacts are shown in column 1 of Table 7; the *p*-values on tests of equality of effects for Black males versus all other males and for Black females versus all other females are 0.004 and 0.010, respectively.).[22] In Table 7 (column 3), we show that Black students similarly drive overall reductions in arrests, with estimated declines of 0.073 and 0.017 arrests for Black male and female students, respectively (the *p*-values on tests of equality are 0.008 for Black males and 0.119 for Black females). 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 percent), above and beyond the identified reduction in OSS days and distinguishable from the estimated absent day impact for all non-Black males (*p*-value of 0.031).

Turning to school climate and academic outcomes, we cannot reject that treatment effects are equal across within-gender racial/ethnic groups (or across gender). Importantly, we do not find evidence that any subgroups are harmed by RP adoption. Estimated school climate impacts are positive across subgroups. Estimated test score impacts for White students are relatively imprecise (consistent with their small sample share), and we cannot reject reading value-added declines up to 0.13 SD (0.07 SD) for White females (males) and math value-added declines up to 0.04 SD (0.07 SD) for White females (males). In contrast, we can reject test score declines larger than 0.03 SD for Black and Latine females and males with only one exception (and we can reject value-added declines greater than 0.01 SD for Black males in both reading and math).

One explanation for the differential arrest, suspension, and absent day declines experienced by Black male students (and the differential decline in out of school suspension days for Black females) is that they may be concentrated in schools that employ RP most effectively. However, in Table A23, we estimate the impact of RP on within-school disparities between Black students and non-Black students in OSS days, arrests and test scores, and we find that declines in OSS days and arrests mirror overall estimates (findings for test scores are quite imprecise and so difficult to interpret). Given that treatment effect heterogeneity on the basis of race and gender persists within schools, our findings are consistent with RP implementation being

[22] See Figures A26–A31 for event studies for each outcome and subgroup. Interestingly, the suspension day declines for Black students exceed the estimated decline (shown in Table 6) 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 RP or may reflect Black student behavior being particularly responsive to RP (Francis 2012).

Table 7—High School Restorative Practices: Race-by-Gender Treatment Heterogeneity

	Out-of-school suspension days (1)	Absent days (2)	Number of arrests (3)	School climate (std) (4)	Reading value added (std) (5)	Math value added (std) (6)
Overall male	−0.147	−0.660	−0.038	0.039	0.014	0.031
	(0.082)	(0.421)	(0.012)	(0.017)	(0.016)	(0.018)
Overall female	−0.156	−0.548	−0.010	0.044	−0.001	0.007
	(0.057)	(0.458)	(0.004)	(0.018)	(0.019)	(0.013)
Test (M = F): *p*-value	0.865	0.746	0.015	0.730	0.577	0.817
Black male	−0.384	−1.655	−0.073	0.041	0.019	0.032
	(0.118)	(0.572)	(0.021)	(0.027)	(0.015)	(0.018)
Test (BM = NBM): *p*-value	0.004	0.031	0.008	0.660	0.817	0.987
Black female	−0.325	−0.658	−0.017	0.061	0.010	0.013
	(0.110)	(0.597)	(0.007)	(0.029)	(0.020)	(0.015)
Test (BF = NBF): *p*-value	0.010	0.591	0.119	0.550	0.242	0.623
Latine male	0.003	−0.051	−0.022	0.017	0.017	0.032
	(0.066)	(0.632)	(0.009)	(0.020)	(0.019)	(0.023)
Test (LM = NLM): *p*-value	0.014	0.099	0.070	0.495	0.783	0.933
Latine female	−0.053	−0.284	−0.004	0.048	−0.006	0.010
	(0.035)	(0.669)	(0.002)	(0.020)	(0.021)	(0.020)
Test (LF = NLF): *p*-value	0.023	0.560	0.132	0.943	0.889	0.623
White male	−0.042	−0.620	0.002	0.041	0.012	0.012
	(0.072)	(0.829)	(0.014)	(0.029)	(0.039)	(0.042)
Test (WM = NWM): *p*-value	0.241	0.903	0.084	0.880	0.836	0.573
White female	−0.075	−0.869	−0.008	0.007	−0.033	0.012
	(0.039)	(0.802)	(0.004)	(0.031)	(0.050)	(0.026)
Test (WF = NWF): *p*-value	0.255	0.674	0.637	0.215	0.374	0.867

Notes: Observations are at the student-school year level and the sample covers students in grades 9 to 12 between SY09 and SY19. See Supplemental 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/ethnicity 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. For each race/ethnicity-by-gender group, we present *p*-values from the test of the null hypothesis that the estimate for that group equals the estimate for all other students of the same gender (i.e., Test (BF = NBF) is the test of the null hypothesis that the treatment effect for Black females equals the treatment effect for non-Black females) Standard errors are clustered at the school level.

relatively homogeneous across schools and with Black students benefiting most in terms of reduced exposure to punitive discipline.

Heterogeneity by English Learner Status, Grade Level, and Disability.—We find significant differences in treatment effects across subgroups only for out-of-school suspension (OSS) days (larger declines among native English speakers and for ninth and tenth graders) and arrests (larger declines for ninth and tenth graders and those classified as having a disability).[23] The patterns we document may reflect the

[23] See Tables A24–A29. Reading test score treatment effects are also larger for those classified as having a disability (the *p*-value on the test of equal effects for students with and without documented disabilities is 0.075), and effects on ISS days differ by English learner status (driven by increases for English learners).

challenges in translating RP to those not fluent in the instructional medium of English (though in the case of arrests, both English learners and native English speakers experience significant declines), the greater scope to adopt new practices and norms at lower grade levels, and/or the higher baseline OSS days and arrests among students at lower grade levels and among those classified as having a disability.

Heterogeneity by RP Program Type.—RP implementation can vary widely, 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 (as described in Section II): RP Coaching, RP Leadership, and RP Peer Council.[24] In Tables A30 and A31, we present RP program-specific estimates and test the null hypothesis that RP Coaching estimates are equal to RP Leadership estimates. While point estimates are consistent with the less-intensive RP Leadership program having failed to contribute to the out-of-school suspension (OSS) and arrest declines we document, associated estimates are imprecise and we cannot reject the null that RP Coaching and RP Leadership treatment effects are equal.

Heterogeneity by Culture of Calm Exposure.—CPS' 'Culture of Calm' Initiative (CoC), launched in SY10, was intended to promote student safety and learning. In theory, prior CoC exposure could increase schools' RP implementation capacity (or mute RP benefits if programmatic features overlap). Estimates of treatment effect heterogeneity by prior exposure to CoC are imprecise, likely due in part to the limited number of schools that received CoC programming.[25] Figure A32 presents event studies that characterize the timing of out-of-school suspension and arrest effects following CoC adoption. While we find an immediate and persistent decline in out-of-school suspension days after CoC exposure, there is no clear reduction in arrests before the rollout of RP begins in event year $t + 3$. In addition to highlighting the direct benefits of CoC exposure (Hinze-Pifer and Sartain 2018), these event studies further clarify that suspension reductions need not lead to contemporaneous arrest declines in our study setting.

VIII. Conclusion

School officials grapple with how to optimally create a safe learning environment. Schools tend to be risk-averse, and the inherently "safe" option is to remove students for any breaches of what is considered to be appropriate conduct. On the other hand, by enforcing a 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.

[24] We assign schools to the first RP program type received. The interpretation of RP Peer Council treatment effects is complicated by the fact that several schools implementing RP Peer Council subsequently implemented RP Coaching.

[25] 47 schools received CoC programming. Of these, 28 subsequently adopted RP during the study period and 19 did not. See Tables A32 and A33 for heterogeneity analyses.

We study the causal impact of the rollout of restorative practices in Chicago Public Schools. Over the course of our study period, 122,000 high school students and 107,000 elementary school students were exposed to RP. Exploiting cross-school variation in the timing of the introduction of RP, we show that RP adoption reduced the number of out-of-school suspension days by 17.8 percent and reduced the number of student arrests by 18.8 percent, with declines in arrests for both violent and nonviolent offenses. We estimate sizable declines in out-of-school arrests and find suggestive evidence of improvements in perceived school climate, indicating that RP adoption is not simply altering how teachers and school administrators respond to behavioral challenges and suggesting that students' experiences in schools improved. Turning to treatment effect heterogeneity, we find that absolute declines in 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. Some practitioners may be concerned that RP benefits students who would otherwise be exposed to punitive discipline while harming their classmates by engendering more permissive behavioral norms. We can rule out average test score value-added declines larger than 0.025 SD, and our results taken as a whole provide some evidence that this tradeoff is not of first-order concern in the study setting (though causal evidence linking contemporaneous disciplinary and academic measures to long-run outcomes of interest is generally lacking). Stepping back, one important caveat is that those schools that were selected to receive RP programming were chosen on the basis of expected gains. As such, our estimates likely represent an upper bound on the average effects of RP adoption that should be anticipated in a broader or less selected sample of schools.

Teachers (and schools) have been found to have important, and varying, effects on behavioral outcomes, beyond test scores, for which we know there are meaningful returns (Jackson 2018; Petek and Pope 2023; Rose, Schellenberg, and Shem-Tov 2022). By sending signals to children about optimal ways to behave and how society should ideally work (Parsons 1959; Dreeben 1967; Bowles and Gintis 1976; Seacrest 2023), school disciplinary policies are similarly expected to reach beyond the creation of conditions for learning in the short term. In particular, exposure to a reparative or restorative approach to addressing behavior may help children to develop the skills (including those related to conflict resolution) needed to more constructively approach challenging situations in life.[26] In future work, we seek to understand how RP exposure shapes students' long-run educational and labor market trajectories as well as their criminal involvement in adulthood.

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): 876–90.

Adukia, Anjali, Benjamin Feigenberg, and Fatemeh Momeni. 2024. "Reparative Ripple Effects? Exploring the Impacts of Sibling Exposure to School-Based Restorative Justice." *AEA Papers and Proceedings* 114: 512–16.

[26] In follow-up work, we find evidence that older sibling exposure to RP influences younger sibling school attendance (Adukia, Feigenberg, and Momeni 2024).

Adukia, Anjali, Benjamin Feigenberg, and Fatemeh Momeni. 2025. *Code for: "From Retributive to Restorative: An Alternative Approach to Justice in Schools*." Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E224141V1.

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.

Anyon, Yolanda, Jeffrey M. Jenson, Inna Altschul, Jordan Farrar, Jeanette McQueen, Eldridge Greer, Barbara Downing, and John Simmons. 2014. "The Persistent Effect of Race and the Promise of Alternatives to Suspension in School Discipline Outcomes." *Children and Youth Services Review* 44: 379–86.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–118.

Augustine, Catherine H., John Engberg, Geoffrey E. Grimm, Emma Lee, Elaine Lin Wang, Karen Christianson, and Andrea A. 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. 2024. "The School-to-Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime." *American Economic Journal: Economic Policy* 16 (4): 165–93.

Baker, Myriam L. 2008. *DPS Restorative Justice Project: Executive Summary, 2007–2008*. Outcomes, Inc.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2): 169–217.

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 H. C. Sant'Anna. 2020. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

Chicago Police Department. 2008–2019. *Chicago Police Department: Administrative Data System*.

Chicago Public Schools. 2009–2019. *Chicago Public Schools: Student Administrative Data System*.

Craig, Ashley C., and David Martin. 2023. "Discipline Reform, School Culture, and Student Achievement." IZA Discussion Paper 15906.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 1688–99.

Dreeben, Robert. 1967. "The Contribution of Schooling to the Learning of Norms." *Harvard Educational Review* 37 (2): 211–37.

Evertson, Carolyn M., and Carol S. Weinstein, eds. 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*. NCJRS, Council of State Governments Justice Center.

Francis, Dania V. 2012. "Sugar and Spice and Everything Nice? Teacher Perceptions of Black Girls in the Classroom." *Review of Black Political Economy* 39 (3): 311–20.

Fulkerson, Andrew. 2001. "The Use of Victim Impact Panels in Domestic Violence Justice Approach." *Contemporary Justice Review* 4 (3–4): 355–68.

Gonzalez, Thalia. 2015. "Socializing Schools: Addressing Racial Disparities in Discipline through Restorative Justice." In *Closing the School Discipline Gap: Equitable Remedies for Excessive Education*, edited by Daniel J. Losen, 151–65. Teachers College Press.

Gonzalez, Thalia. 2012. "Keeping Kids in Schools: Restorative Justice, Punitive Discipline, and the School to Prison Pipeline." *Journal of Law and Education* 41 (2): 281–335.

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): 167–82.

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

Hart, Holly, Christopher Young, Alicia Chen, Naureen Kheraj, and Elaine M. Allensworth. 2021. *5Essentials: Evidence-Based School Improvement: School Improvement and School Climate in High Poverty Schools*. UChicago Consortium.

Hashim, Ayesha K., Katharine O. Strunk, and Tasminda K. Dhaliwal. 2018. "Justice for All? Suspension Bans and Restorative Justice Programs in the Los Angeles Unified School District." *Peabody Journal of Education* 93 (2): 174–89.

Hinze-Pifer, Rebecca, and Lauren Sartain. 2018. "Rethinking Universal Suspension for Severe Student Behavior." *Peabody Journal of Education* 93 (2): 228–43.

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): 2072–107.

Karp, David R., and Beau Breslin. 2001. "Restorative Justice in School Communities." *Youth and Society* 33 (2): 249–72.

Kauffman, James M., Patricia L. Pullen, Mark P. Mostert, and Stanley C. Trent. 2011. *Managing Classroom Behavior: A Reflective Case-Based Approach*. Pearson.

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.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher* 49 (4): 241–53.

Lai, Ijun. 2018. "Short-Term Impacts of Chicago's Suspensions and Expulsions Reduction Plan (SERP)." PhD diss., Michigan State University.

Levenstein, Rachel, Sue Sporte, and Elaine Allensworth. 2011. *Findings from an Investigation into the Culture of Calm Initiative*. UChicago Consortium.

Lewis, Sharon. 2009. *Improving School Climate: Findings from Schools Implementing Restorative Practices*. International Institute for Restorative Practices.

Losen, Daniel, Damon Hewitt, and Ivory Toldson. 2014. *Eliminating Excessive and Unfair Exclusionary Discipline in Schools Policy Recommendations for Reducing Disparities*. Equity Project at Indiana University.

Losen, Daniel J., Cheri L. Hodson, Michael A. Keith II, Katrina Morrison, and Shakti Belway. 2015. *Are We Closing the School Discipline Gap?* Center for Civil Rights Remedies.

Losen, Daniel J., and Paul Martinez. 2020. *Lost Opportunities: How Disparate School Discipline Continues to Drive Differences in the Opportunity to Learn*. Learning Policy Institute.

McCold, Paul. 2008. "Evaluation of a Restorative Milieu: Restorative Practices in Context." In *Restorative Justice: From Theory to Practice*, edited by Holly Ventura Miller, 99–137. Emerald Group Publishing.

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.

McMorris, Barbara J., Kara J. Beckman, Glynis Shea, Jenna Baumgartner, and Rachel C. Eggert. 2013. "Applying Restorative Justice Practices to Minneapolis Public Schools Students Recommended for Possible Expulsion."*Applied Clinical Informatics* 4 (3): 434–44.

Minow, Martha. 2022. "Restorative Justice and Anti-Racism." *Nevada Law Journal* 22 (3): 1157–78.

Mirsky, Laura. 2007. "SaferSanerSchools: Transforming School Cultures with Restorative Practices." *Reclaiming Children and Youth* 16 (2): 5–12.

Parsons, Talcott. 1959. "The School Class as a Social System: Some of Its Functions in American Society." *Harvard Educational Review* 29 (4): 297–318.

Petek, Nathan, and Nolan G. Pope. 2023. "The Multidimensional Impact of Teachers on Students." *Journal of Political Economy* 131 (4): 1057–107.

Pope, Nolan, and George Zuo. 2020. "Suspending Suspensions: The Education Production Consequences of School Suspension Policies." Unpublished.

Public Agenda Foundation. 2004. *Teaching Interrupted. Do Discipline Policies in Today's Public Schools Foster the Common Good?* Public Agenda Foundation.

Rawe, Julie. 2023. "The Difference between IEPs and 504 Plans." *Understood*. https://www.understood.org/en/articles/the-difference-between-ieps-and-504-plans.

Riestenberg, Nancy. 2003. "Restorative Schools Grants Final Report, January 2002–June 2003: A Summary of the Grantees' Evaluation." Unpublished.

Rose, Evan K., Jonathan T. Schellenberg, and Yotam Shem-Tov. 2022. "The Effects of Teacher Quality on Adult Criminal Justice Contact." NBER Working Paper 30274.

Ross, Lee, and Constance Stillinger. 1991. "Barriers to Conflict Resolution." *Negotiation Journal* 7 (4): 389–404.

Sartain, Lauren, Elaine M. Allensworth, Shanette Porter, Rachel Levenstein, David W. Johnson, Michelle Hanh Huynh, Eleanor Anderson, Nick Mader, and Matthew P. Steinberg. 2015. *Suspending Chicago's Students: Difference in Discipline Practices across Schools*. UChicago Consortium.

Seacrest, Logan. 2023. *Justice for All: How Restorative Justice Mutually Benefits Victims and Youth*. R. Street.

Shem-Tov, Yotam, Steven Raphael, and Alissa Skog. 2024. "Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program." *Econometrica* 92 (1): 61–78.

Shollenberger, Tracey L. 2015. "Racial Disparities in School Suspension and Subsequent Outcomes: Evidence from the National Longitudinal Survey of Youth." In *Closing the School Discipline Gap: Equitable Remedies for Excessive Exclusion*, edited by Daniel J. Losen, 31–44. Teachers College Press.

Stevens, W. David, Lauren Sartain, Elaine M. Allensworth, Rachel Levenstein, Shannon Guiltinan, Nick Mader, Michelle Hanh Huynh, and Shanette Porter. 2015. *Discipline Practices in Chicago Schools: Trends in the Use of Suspensions and Arrests*. UChicago Consortium.

Strang, Heather, Lawrence W. Sherman, Evan Mayo-Wilso, Daniel Woods, and Barak Ariel. 2013. "Restorative Justice Conferencing (RJC) Using Face-to-Face Meetings of Offenders and Victims: Effects on Offender Recidivism and Victim Satisfaction. A Systematic Review." *Campbell Systematic Reviews* 9 (1): 1–59.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.

Teasley, Martell L. 2014. "Shifting from Zero Tolerance to Restorative Justice in Schools." *Children and Schools* 36 (3): 131–33.

Twemlow, Stuart W., Peter Fonagy, and Frank C. Sacco. 2004. "The Role of the Bystander in the Social Architecture of Bullying and Violence in Schools and Communities." *Annals of the New York Academy of Sciences* 1036 (1): 215–32.

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.

Wilson-Simmons, Renée, Kimberly Dash, Parisa Tehranifar, Lydia O'Donnell, and Ann Stueve. 2006. "What Can Student Bystanders Do to Prevent School Violence? Perceptions of Students and School Staff." *Journal of School Violence* 5 (1): 43–62.

Winn, Maisha. 2016. "Transforming Justice: Transforming Teacher Education." Unpublished.

Wolf, Kerrin C., and Aaron Kupchik. 2017. "School Suspensions and Adverse Experiences in Adulthood." *Justice Quarterly* 34 (3): 407–30.

Zagar, Robert John, Joseph W. Kovach, Terry Ferrari, William M. Grove, Kenneth G. Busch, John Russell Hughes, and Agata Karolina Zagar. 2013. "Applying Best Treatments by Using a Regression Equation to Target Violence-Prone Youth: A Review." *Comprehensive Psychology* 2: 03.16.49. CP.2.7.