ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-442

From Referrals to Suspensions: New Evidence on Racial Disparities in Exclusionary Discipline

Jing Liu
University of Maryland
College Park

Michael S. Hayes
Rutgers University - Camden

Seth Gershenson
American University and IZA

We use novel data on disciplinary referrals, including those that do not lead to suspensions, to better understand the origins of racial disparities in exclusionary discipline. We find significant differences between Black and white students in both referral rates and the rate at which referrals convert to suspensions. An infraction fixed-effects research design that compares the disciplinary outcomes of white and non-white students who were involved in the same multi-student incident identifies systematic racial biases in sentencing decisions. On both the intensive and extensive margins, Black and Hispanic students receive harsher sentences than their white co-conspirators. This result is driven by high school infractions and mainly applies to "more severe" infractions that involve fights or drugs. Reducing racial disparities in exclusionary discipline will require addressing underlying gaps in disciplinary referrals and the systematic biases that appear in the adjudication process.

VERSION: July 2021

Suggested citation: Liu, Jing, Michael S. Hayes, and Seth Gershenson. (2022). From Referrals to Suspensions: New Evidence on Racial Disparities in Exclusionary Discipline. (EdWorkingPaper: 21-442). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/axvg-zp19

From Referrals to Suspensions: New Evidence on Racial Disparities in Exclusionary Discipline *

Jing Liu[†]

Michael S. Hayes[‡]

Seth Gershenson[§]

Abstract: We use novel data on disciplinary referrals, including those that do not lead to suspensions, to better understand the origins of racial disparities in exclusionary discipline. We find significant differences between Black and white students in both referral rates and the rate at which referrals convert to suspensions. An infraction fixed-effects research design that compares the disciplinary outcomes of white and non-white students who were involved in the same multi-student incident identifies systematic racial biases in sentencing decisions. On both the intensive and extensive margins, Black and Hispanic students receive harsher sentences than their white co-conspirators. This result is driven by high school infractions and mainly applies to "more severe" infractions that involve fights or drugs. Reducing racial disparities in exclusionary discipline will require addressing underlying gaps in disciplinary referrals and the systematic biases that appear in the adjudication process.

Keywords: Exclusionary discipline, intentional discrimination, office referrals
JEL Classification: I2, J7

*The authors thank participants in AEFP, APPAM, and SOLE, and seminars at George Mason University, Indiana University, New York University, and University of Maryland, for helpful discussions. Wenjing Gao provides excellent research assistance. The data were provided via in-kind support from a local educational agency using a data sharing agreement that requires agency review of the findings of the research prior to publication. Opinions reflect those of the authors and not necessarily those of the funding agencies or the data-sharing partner.

[†][Corresponding author] jliu28@umd.edu. University of Maryland College Park and IZA.

[‡]michael.hayes@rutgers.edu. Rutgers University – Camden.

[§]gershens@american.edu. American University and IZA.

1 Introduction

Racial disparities in exclusionary discipline (i.e., suspensions) in U.S. public schools are striking: for example, the 2013-14 Civil Rights Data Collection finds that Black students accounted for 40% of suspensions but only 16% of enrollments. These discrepancies frequently arise between students in the same school or district, particularly in large, integrated urban school districts (Chin, 2021). Such disparities are the subject of much debate and concern for two broad reasons. First, suspensions likely affect important socioeconomic outcomes and are thus a precursor to analogous disparities in educational achievement, high school and college completion, employment, and involvement with the criminal justice system (Bacher-Hicks et al., 2019; Davison et al., 2021; Sorensen et al., Forthcoming; Weisburst, 2019), all of which create an array of social costs that cities must absorb. This motivates efforts to reduce the use of exclusionary discipline, which disproportionately harms students of color (Steinberg and Lacoe, 2017; Davison et al., 2021). Second, racial disparities in exclusionary discipline may be artificial in the sense that they result from systematic biases in schools' handling of student indiscipline and not underlying racial differences in student behavior.

The prevalence of these biases, or "intentional discrimination" as per a 2014 Dear Colleague Letter from the Obama Administration, has implications for how schools and policy makers might go about reducing racial disparities in exclusionary discipline and for reducing its use more broadly. Doing so is important, as racial disparities in educational outcomes, which are indicative of untapped potential, suggest that many cities are missing out on the myriad benefits of a well educated citizenry. For example, higher levels of education are associated with increased productivity (Moretti, 2004), more civic engagement (Dee, 2004), and reduced crime (Lochner and Moretti, 2004). Moreover, there are budgetary implications for cities in terms of lost tax revenue and increased social spending (Sum et al., 2009).

Causal identification of systematic biases in sentencing decisions is challenging because no two infractions are identical, and researchers typically do not observe the student behaviors

that lead to student suspensions.[1] Barrett et al. (2019) introduce a novel solution: compare the suspension lengths (in days) received by students of different races who were involved in the *same* incident.[2] Using administrative data on suspensions in Louisiana, the authors use an incident fixed effects (FE) strategy to compare student-specific disciplinary outcomes following fights between Black and white students. They find that Black students receive longer suspensions, on average, than their white counterparts. The difference is modest in size but statistically significant. This finding suggests that intentional discrimination in the adjudication of these fights contributes to the Black-white suspension gap.

We extend this approach to testing for intentional discrimination and probe the identifying assumptions in a few ways using rich administrative data from a large and diverse urban school district in California with sizable enrollments of white, Black, Hispanic, and Asian students. A primary contribution of our study is to rigorously test for racial bias in exclusionary discipline in a new context outside the American South, particularly among Hispanic students, who constitute the fastest growing ethnic group in the U.S.

Our second contribution relies on disciplinary referral data in addition to realized suspensions, as not all referrals lead to a suspension. This is important for a few reasons. First, if suspensions are the sole measure of misbehavior, prior referrals that did not lead to a suspension are an omitted variable that could influence the suspension assigned to subsequent incidents.[3] Second, by relying solely on suspensions, Barrett et al. (2019) omit students

[1] An analogous challenge exists in studies of racial bias in police's use of force (Fryer Jr, 2019).

[2] Shi and Zhu (2021) adopt a similar strategy and replicate these findings in North Carolina.

[3] A simple example illustrates the problem: 1) suppose two students, one Black and one white, are otherwise identical in terms of socioeconomic and academic background, and are in the same classes; 2) they participate equally in a fight, and receive suspensions of 5 and 2 days, respectively 3) this was the first suspension of the school year for each student. This is the data available in previous research (e.g., Barret et al. 2019), and from this information it looks like a clear case of intentional discrimination, as the Black student received a harsher punishment than the white student, even though they had "identical" backgrounds and participated equally in a singular disciplinary incident. However, now consider some additional information: 1) the principal's leniency decreases with each incident (referral); 2) this was the Black student's third disciplinary referral but the white student's first. Assuming that there is no systemic racial bias in the office referrals themselves (i.e., they're accurate), this new information makes the difference in suspension length seems less arbitrary, less biased, and more the result of underlying referral histories.

Whether racial bias exists in office referrals is an empirical question, of course, which falls outside the scope of the current study. It likely does, since teachers are the primary referrers and teachers' assessments

who were involved in the same fight but were not suspended. This form of sampling on the dependent variable is potentially problematic because there are consequences on the extensive margin (being suspended) over and above those of being suspended for an additional day, which would under-estimate the magnitude of intentional discrimination. Observing all referrals in addition to associated suspension outcomes allows us to avoid both problems.

Third, the universe of referrals allows us to test for intentional discrimination in all types of disciplinary infractions and not just fights. This is useful because fights are potentially unique in terms of having an instigator or a "more violent" participant, which might lead to an omitted variables bias, and because principals' biases might vary by infraction type. Moreover, knowing whether intentional discrimination is more pronounced for certain types of infractions provides critical information for the design of interventions and policies that aim to reduce racial disparities in exclusionary discipline.

Finally, these data facilitate one of the first systematic, quantitative descriptions of the referral process, as nearly all existing research on racial disparities in exclusionary discipline focuses on suspensions (e.g., Anderson et al. 2017; Bacher-Hicks et al. 2019; Barrett et al. 2019; Holt and Gershenson 2019; Lindsay and Hart 2017; Kinsler 2011) and not the referral and reporting process that necessarily precedes the decision of whether, and for how long, to suspend a student.[4] Referrals merit the attention of researchers and policymakers independent of their connection to suspensions because even when referrals do not result in exclusionary discipline, they are intermediate educational outcomes that can erode students' trust in teachers, the quality of student-teacher relationships, and students' engagement in school. In turn, strained student-teacher relationships and student disengagement can harm achievement and lead to future disciplinary infractions. While we cannot distinguish

of student behavior are known to be biased (Dee, 2005). That said, racial biases against students of color in the referral process would not explain the sentencing disparity provided in the current example; if anything, principals aware of this might be more lenient with students of color, such that the discrepancies we observe form a lower bound of sorts for the true amount of intentional discrimination in the sentencing process. In any case, relying on suspension data alone means that it is impossible to know, and thus to account for, a student's full disciplinary history, which in turn can lead to misdiagnoses of intentional discrimination.

[4]An exception is Girvan et al. (2017), who conduct descriptive analyses of referral (but not suspension) data and conclude that implicit bias among teachers contribute to racial gaps in office referrals.

racial disparities in referrals from underlying behavioral differences between racial groups, the current study advances our understanding of the role the referral process plays in racial disparities in exclusionary discipline.

We begin our analyses by describing the distribution of disciplinary referrals and the rate at which referrals result in suspensions. Decompositions of the large, unconditional Black-white gaps in both suspensions and referrals show that these gaps are primarily driven by within-school variation. For example, Black students are about 4 percentage points more likely to have been suspended in a given year than their white peers in the same school. However, we go beyond past research on suspensions by conducting similar analyses of disciplinary referrals and find that Black students are 12 percentage points more likely to have received at least one disciplinary referral than their white peers in the same school. This suggests that part of the racial gap in suspensions is due to underlying differences in the frequency of office referrals. However, the racial gap in referral propensities is not the sole reason for the racial gap in suspensions, as we also find that the conversion rate of referrals into suspensions is significantly higher for Black than for white students.

Following Barrett et al. (2019), we then test for intentional discrimination by using an infraction-FE approach. These estimates show a clear and consistent pattern in which Black and Hispanic students are punished more severely than white students who were involved in the same incident and had the same prior disciplinary histories. Specifically, Black students were about 2 percentage points (67%) more likely to be suspended than white students involved in the exact same incident. This finding is robust to controlling for past achievement, referrals, and suspensions, suggesting that intentional discrimination explains a nontrivial share of this disparity. Interestingly, this type of intentional discrimination seems confined to high schools and more severe types of incidents.

2 Data

Administrative data come from a large and demographically diverse urban school district in California for the 2016-17 through 2019-20 school years. Panel A of Table 1 summarizes the student-by-year level analytic sample. The district served 84,056 unique students in grades K-12 (240,652 student-year observations in about 200 unique schools each year) during this time, of which 12% are white, 7% are Black, 30% are Hispanic, and 33% are Asian. We use students' home addresses to identify their residential census tract, which we then use to create a proxy for students' socioeconomic status (i.e., poverty rate in their neighborhood).

The distinguishing feature of the data is detailed information on disciplinary referrals, regardless of whether they lead to a suspension. Specifically, referral records include the individual who made the referral, the reason for the referral (i.e., type of incident), and the exact time, date, and location of the incident (e.g., 3pm, in the hallway, on Monday April 2nd). This precise information allows us to identify the multi-student incidents that are central to our main identification strategy. There were 78,127 unique incidents, of which 12.2% (9,562) involved multiple students. 54.3% of those involved students of different races, which provide identifying variation for the incident-FE identification strategy. The data also uniquely link referrals to suspension outcomes (measured in days).

A few caveats of the referral data are warranted, which are largely analogous to challenges associated with studying citizen-police interactions (e.g., (Fryer Jr, 2019)).[5] Namely, there could be systematic racial biases in who receives a referral; pre-referral behavior is unobserved by the econometrician, so there is no straightforward way to test for this. Such biases could affect both whether a student is referred at all and the severity/category of referrals that do get made. For example, consciously or not, teachers might spend less time monitoring white students or be more prone to downplaying, re-classifying, or outright excusing white students' misbehavior. This complicates comparing referral rates across racial groups, as

[5]One way in which school discipline is easier to study than police interactions is that the "risk set" is clearly defined as all students in the school or classroom.

the mapping from misbehavior to referrals could vary by race. It also calls into question the interpretation of analyses of suspension outcomes that condition on referrals (or referral type). A rigorous analysis of the determinants of referrals, including teachers' biases, is outside the scope of the current study, though we consider the potential implications of these biases when discussing the results.

Panel B of Table 1 summarizes disciplinary outcomes at the student-year level. Column 1 shows that each year about 8% of students received at least one office referral. Among those who had at least one referral, the average student was referred about 4.6 times. These frequencies are higher than for suspensions, indicating that many referrals do not lead to a suspension: only 2% of students were suspended per year and among those suspended, the average student was suspended about 1.6 times for about 3.2 days. We measure the "conversion rate" as the ratio of suspensions to referrals, which is about 5% on average.

Columns 2-6 report these figures separately by the mutually exclusive race/ethnicity categories contained in the administrative data. The "Other" category contains multi-racial, American Indian, Arabic, Samoan, and other non-white students. Comparing across columns, we see stark and statistically significant disparities on both the intensive and extensive margins in both referrals and suspensions. These gaps are largest when comparing Black students to white and Asian students: Black students are more than 5 times as likely to be referred and 7 times more likely to be suspended in a given year than white students, for example. There is a smaller but still sizable white-Hispanic gap as well. Appendix Table A1 reports referral rates by student race and school type. Referrals are most common in middle schools, in both absolute and relative terms, though they occur in all grade levels.

The data also provide the reason(s) for each referral. Many referrals are the result of multiple infractions, so for the purpose of heterogeneity analyses we follow Lindsay and Hart (2017) in making mutually exclusive, one-off categories based on the "most severe" reason listed for the referral: (a) violence; (b) drugs; (c) interpersonal offenses; (d) disruption or noncompliance; (e) class skipping or walkout; (f) other. For example, a referral where the

student was charged with both class skipping and disruption would be coded as disruption. Appendix Table A2 summarizes the types of referrals by school type. Different types of referrals occur at different rates across school types, as might be expected. For example, drug-related offenses are rare overall, but predominantly occur in high school. Interpersonal offenses and offenses due to disruption, noncompliance, class skipping, or walkout are more prevalent in middle schools. Violence incidents are most common in elementary school.

3 Methods

We begin the descriptive analysis by decomposing raw Black-white and Hispanic-white referral and suspension gaps. To do this, we first calculate the weighted average for a given disciplinary outcome for each racial group and grade across all schools, and the weights are the number of students in a school-grade-race cell. We then derive the overall racial gaps using these averages and decompose them into between- and within-school gaps [6] We then further drill down into within-school gaps by estimating linear regressions at the *student-year* level that condition on a host of student characteristics and FEs. The main outcomes for these regressions are indicators for ever referred and ever suspended in a given year. To examine the intensive margin, we consider outcomes including total referrals, total suspensions, and the likelihood that a referral results in a suspension. Regressions for these outcomes are estimated on the restricted sample of students who had at least one referral in a year.[7]

Specifically, we estimate models of the form

$$Y_{ist} = \beta Race_i + \gamma X_{ist} + \theta_{st} + \epsilon_{ist}, \tag{1}$$

[6] This exercise follows Barrett et al. (2019) and Clotfelter et al. (2005); see Appendix B for details.

[7] As discussed in section 2, these regressions are limited in the sense that low-level infractions might be ignored for certain students, such that receiving a referral could mean different things for different students. However, if this referral bias works in favor of white students, and we see that conditional on receiving a referral students of color receive longer suspensions on average, then this is likely an underestimate of the degree to which students of color are punished more harshly than their white classmates.

where Y_{ist} is a disciplinary outcome for student i in school s in year t. We estimate Equation (1) with and without covariates (X_{ist}), where X includes lagged academic achievement and discipline outcomes, gender, neighborhood poverty rates, and special education status. This descriptive exercise provides novel, suggestive evidence that racial gaps in referrals *and* in the processing of referrals contribute to racial gaps in suspensions.

However, statistically significant estimates of β do not necessarily indicate the presence of racial bias, as these models do not control for the severity or frequency of the infractions that led to the referral. Following Barrett et al. (2019), we address this omitted variables concern by switching to *student-by-incident* level analyses and comparing suspension outcomes for students of different races who were involved in the same incident. This is an unambiguous improvement over controlling for incident type, since no two incidents are identical, and there could be racial biases in how incident type is coded. The latter remains a concern in terms of sample construction and external validity, which we discuss in more detail below. Importantly, these analyses include students who were not suspended at all, as incidents are defined by referrals and not suspensions.

Specifically, we estimate models of the form

$$S_{ijt} = \alpha Discipline_{i,j-1,t} + \beta Race_i + \gamma X_{it} + \theta_j + \epsilon_{ijt}, \tag{2}$$

where S is the suspension outcome (in days or an indicator for suspension) awarded to student i stemming from incident j. Similar to Equation (1), we control for prior year's test scores and disciplinary incidents. Because we are using incident-level data, we can further control for student i's disciplinary incidents in the current year that occurred prior to incident j in Equation (2) to account for the possibility that principals consider the student's entire history of referrals before making a decision. Most importantly, we use incident FE (θ_j) to exploit within-infraction variation in disciplinary outcomes. These FEs control for unobserved aspects of the severity and nature of the incident and make school

and year FE redundant, as incidents can only involve students in the same school.

There are two threats to the validity of OLS estimates of β in Equation (2). The first regards internal validity. The concern here is that, on average, students of different races did not participate "equally" in the incident in terms of instigation, showing remorse, or degree of misbehavior. For instance, if white students were more likely to be the instigator of fights, then comparing the disciplinary outcomes of white and non-white students who fought will conflate intentional discrimination with a harsher penalty for instigation. We cannot directly rule out this possibility, but we probe this question by applying the same model to different types of incidents. Intuitively, some incidents, like drugs or class skipping, are less likely to have an instigator or "heavier" participant, and thus the FE estimates provide an arguably more valid comparison.

The second threat regards external validity, as identification comes from a selected sample of multi-student, multi-race incidents (Miller et al., 2019). Appendix Table A3 shows that the identifying sample differs systematically from the overall sample in terms of incident size, racial composition, and prior achievement. This means incident-FE estimates may not generalize to the full population. Moreover, if treatment effects are heterogeneous, the estimates can be biased. For example, if drug incidents are more likely to involve multiple students from different racial/ethnic backgrounds, and Black students are punished more harshly than white students in these incidents, our estimates would be biased upwards. Following Miller et al. (2019), we conduct a weighting exercise based on the predicted likelihood of being in the identifying sample to verify that our findings are robust to this threat.

Finally, if teachers' biases in referrals lead to multi-race incidents not being identified as such because the white participant was not referred, and this is more likely to occur for infractions involving unequal participation between students, this is actually good because unequal participation is a fundamental thread to validity. However, if this sort of non-reporting is more common for less serious offenses, it will merely reduce the generalizability of our results and our ability to precisely identify heterogeneous effects by incident type.

4 Main Results

4.1 Decomposing Racial Gaps in Referrals and Suspensions

Figure 1 decomposes racial gaps in referral and suspension rates into between- and within-school components separately by grade. Panel A shows a sizable Black-white referral rate gap of 10 to 30 percentage points in each grade, which peaks in middle school. Two-thirds of the gap is due to within-school differences, suggesting that the gap is not due to racial sorting into schools. Panel B shows similar patterns in Black-white suspension gaps that are consistent with those in Louisiana (Barrett et al., 2019).

Panels C and D report the same figures for Hispanic-white gaps, which tend to be smaller but otherwise similar to analogous Black-white gaps. A key difference, however, is that the Hispanic-white suspension rate gap is much smaller and more closely resembles the Asian-white gap shown in panel F. The other notable difference is that overall, between-school differences tend to play a larger role in explaining the Hispanic-white gap.

The Asian-white gaps summarized in panels E and F are close to zero in the early grades and slightly favor Asian students in high school. Interestingly, the only case of the between- and within-school gaps diverging is for the Asian-white referral gap shown in panel E, where the between-school gap favors white students and the within-school gap favors Asian students. This suggests unique school sorting patterns for Asian students.

4.2 Racial Gaps in Referrals, Suspensions, and Conversions

The decomposition exercises reported in Figure 1 show that racial differences in referral and suspension rates are not merely a product of sorting into schools, as nontrivial shares of these gaps are driven by within-school differences. However, even within-schools racial differences in students' behavior or backgrounds could explain the differences. Table 2 reports estimates

of Equation (1) that control for school-by-year FE and time-varying student covariates.[8]

Each column of Table 2 reports regression-adjusted racial gaps in a specific disciplinary outcome (relative to white students). Columns 1 and 2 report estimates for the extensive margin of receiving at least one suspension and at least one referral, respectively. Consistent with prior research, Black students are more likely to be suspended than white students. Specifically, the likelihood of receiving at least one suspension for a typical Black student is 3.2 percentage points higher than for a white student in the same school with the same observed academic and disciplinary history. Column 2 shows an even larger Black-white gap in the chances of receiving a referral of about 10.1 percentage points. This suggests that the disparity in referrals contributes to the gap in suspensions. These point estimates are six and two times larger than baseline (white student) suspension and referral rates of 0.7% and 4.6%, respectively. Hispanic-white gaps are smaller in both absolute and relative terms than the Black-white gap, although both are at least marginally statistically significant. Sizable and statistically significant Asian-white gaps favor Asian students.

Columns 3 and 4 of Table 2 report estimates of racial gaps on the intensive margin of total annual referrals and suspensions. On average, Black students received 0.14 more suspensions than white students. In contrast, as shown in Column 4, the typical Black student received 2.02 more office referrals than the typical white student. Analogous Hispanic-white gaps for these two outcomes are not significantly different from zero after adjusting for covariates, while a modest Asian-white gap in favor of Asian students remains.

The Black-white gap in referrals is an order of magnitude larger than the analogous gap in suspensions, which suggests that there are racial differences in the rate at which referrals convert to suspensions. Column 5 confirms this by estimating models in which the outcome is the ratio of each student's suspensions to referrals. Conditional on student demographics and prior discipline history, conversion rates are similar for white, Asian, and

[8]A parsimonious specification with FE but no student-level controls, reported in Appendix Table A4, provides qualitatively similar results that suggest racial disparities are not due to observable differences in students' backgrounds. The results are also robust to controlling for principal FE or principal tenure.

Hispanic students. However, the conversion rate for Black students is significantly greater (1.8 percentage points, or 47%) than for any other group. Together, the results in Table 2 suggest that Black-white gaps in suspensions are due to disparities in the frequency of disciplinary referrals *and* in the rate at which referrals convert to suspensions.

4.3 Intentional Discrimination

Table 3 reports baseline estimates of Equation (2), which compare the disciplinary outcomes of students involved in the same multi-student infraction. The unit of analysis is the student-incident, so students who were involved in multiple multi-student events appear in the data multiple times. To preserve power, in panel A of Table 3, we group Black, Hispanic, and "other" students together, as these are the groups most at risk of receiving exclusionary discipline.[9] Panel B of Table 3 re-estimates the model with a full set of race/ethnicity indicators. We focus on whether a student was suspended at least once in Table 3, which is arguably the most policy relevant outcome, though we report analogous results in Table A5 in which the outcome is the length of all suspensions (in days and including zeros).

Column (1) of Table 3 reports estimates of a simple model that only controls for infraction fixed effects. Subsequent columns add additional controls, including student characteristics, lagged test scores, lagged disciplinary incidents, and finally current-year incidents that occurred prior to the current incident. The estimated coefficients on the race/ethnicity indicators are qualitatively similar across model specifications, suggesting that infraction fixed effects do a decent job of controlling for possibly confounding factors. However, the estimates in column (3) are slightly smaller and more precisely estimated, which suggests that it is important to control for previous referrals and suspensions in the current year. Accordingly, the fully specified regressions in Column (3) are our preferred estimates, which if anything might be slightly attenuated by the inclusion of current year referrals.

[9]The other category includes many mixed-race students and generally students in this category resemble Black and Hispanic students in terms of other observable characteristics.

Column (3) of Panel A shows that on average, Black and Hispanic students were 2 percentage points more likely to be suspended than their white same-incident peers. Relative to the baseline white suspension rate in the analytic sample, this indicates a large (69%) increase in the likelihood of suspension and provides strong evidence of systematic bias in adjudications. Asian students were slightly more likely to be suspended than white same-incident peers, but this difference is not statistically significant at traditional confidence levels. The analogous estimates in Column (3) of Panel B show that systematic bias is not unique to one group, but roughly similar across Black, Hispanic, and "other-race" students.

The results for suspension length in Table A5 are qualitatively similar to those in Table 3. Once again, the point estimates are robust. Panel B shows that gaps are largest for Black students, but qualitatively similar to those of Hispanic students. Following Barrett et al. (2019), in Appendix Table A6 we restrict the analytic sample to incidents that were each student's first of the year and find qualitatively similar, yet less precise, estimates. The findings presented in Tables 3 and A5 are also robust to implementing the weighting procedure suggested by Miller et al. (2019); see Appendix Table A7).[10]

Having documented systematic racial bias in the district's disciplinary adjudications, we now test for heterogeneity along several dimensions to understand where these biases are most pronounced. The descriptive analysis in section 4.1 shows that raw racial gaps in referrals and suspensions peak in middle school and that the most common reasons for referrals vary by grade level. This suggests that racial biases might also vary by grade level. Accordingly, we re-estimate the preferred full-specification of Equation (2) separately by school type (i.e., elementary, middle, and high) in Appendix Table A8. Here, we see that racial biases in adjudications are almost entirely driven by decisions made for high school students. Interestingly, we also find similar levels of intentional discrimination against Black and Hispanic high school students. Finally, and somewhat surprisingly, we find substantial

[10]See Appendix C for details. Miller et al. (2019)'s discussion is limited to binary treatment status. Since we have multiple student racial groups, for simplicity, these weighted regressions adopt a binary "treatment" where white and Asian students are compared to disproportionately suspended Black, Hispanic, and "other" students. Point estimates are slightly smaller, but remain statistically significant.

discrimination against Asian students in elementary school. This finding merits further consideration, though could be driven by outliers in the relatively small number of multi-student, multi-race incidents in elementary school that involve an Asian student.

One possible interpretation of the finding that intentional discrimination is most prevalent in high schools is that certain types of offenses, which predominantly occur in high schools, are more susceptible to subjective interpretations that lead to biased punishments (e.g., defiance). To investigate, in Table 4 we re-estimate the preferred model separately by referral reason. We see clear heterogeneity by incident type. Violence is the only type of referral that yields statistically significant differences between white and nonwhite students, with magnitudes that are remarkably similar across all non-white demographic groups. Drug incidents yield larger, but imprecisely estimated gaps. Estimated disparities get even smaller and approach zero for less severe incidents such as defiance and class-skipping. Overall, these results suggest that racial discrimination seems to be most salient in severe incidents involving fighting or violence. Of course, keep in mind that the underlying behavior is unobserved, and this result could be due to either students of color being systematically more violent in multi-student incidents, or to there being analogous bias in the content of referrals themselves. The former is arguably unlikely, especially for observably similar students in the same school with similar behavior histories, though biases in the referral process could certainly play a role.

Finally, in Appendix Table A9 we test for heterogeneity in intentional discrimination by gender, special education status, and neighborhood poverty level, as there are differences between these groups in exclusionary discipline (Mendez and Knoff, 2003; Steinberg and Lacoe, 2017); however, we find no evidence of heterogeneity along these dimensions.

5 Conclusion

This study investigates two potential sources of racial disparities in exclusionary discipline (suspensions). First, the gap could be the natural result of analogous disparities in dis-

ciplinary referrals. Second, there could be systematic biases in the adjudication of such referrals. We find evidence that both explanations likely contribute to large and troubling racial disparities in exclusionary discipline. Specifically, using unusually detailed administrative data from a large and diverse urban school district in California, we show that Black-white disparities in exclusionary discipline are large, present in all grade levels, largest in middle school, and primarily due to within- rather than between-school differences. We expand on this descriptive result, which has been documented elsewhere, by showing that similar patterns exist in disciplinary referrals. Finally, we expand on Barrett et al. (2019) to test for systematic racial biases in the adjudication of referrals. Importantly, the referral data allow us to include students who do not get suspended, to fully control for students' prior discipline histories, and to include for the full range of incident types beyond just fights.

We find suggestive evidence of systematic racial bias in the district's disciplinary adjudications. Specifically, compared to white students involved in the same incident who had similar prior disciplinary histories, on average, Black and Hispanic students were 67% (2 percentage points) more likely to be suspended. These estimates are in line with results in Barrett et al. (2019) and Shi and Zhu (2021), which is itself a striking result given that the studied context, a large urban district in California, is at the forefront of efforts to address racial inequities in exclusionary discipline. Heterogeneity analyses show that racial biases in adjudications mainly occur in high school and for violent infractions (e.g., fights).

Closing racial gaps in exclusionary discipline therefore requires addressing both gaps in referrals and biases in the adjudication process. However, there are several issues the current study does not speak to, most importantly whether there are similar biases in the referral process. An implication for schools may be to leverage insights from social psychology regarding empathy interventions, which have been shown to change teachers' perceptions, reduce suspensions, and improve students' achievement (Okonofua et al., 2016). Future research should work to understand the types of teachers, school personnel, and schools that generate these disparities, and the conditions in which they do so.

References

Anderson, Kaitlin, Gary Ritter, and Gema Zamarro, "Understanding a vicious cycle: Do out-of-school suspensions impact student test scores?," 2017.

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

Barrett, Nathan, Andrew McEachin, Jonathan N Mills, and Jon Valant, "Disparities and discrimination in student discipline by race and family income," *Journal of Human Resources*, 2019, pp. 0118–9267R2.

Chin, Mark J, "Desegregated but still separated? The impact of school integration on student suspensions and special education classification," *Journal of Urban Economics*, 2021.

Clotfelter, Charles T, Helen F Ladd, and Jacob Vigdor, "Who teaches whom? Race and the distribution of novice teachers," *Economics of Education review*, 2005, *24* (4), 377–392.

Davison, Miles, Andrew Penner, Emily Penner, Nikolas Pharris-Ciurej, Sonya R. Porter, Evan Rose, Yotam Shem-Tov, and Paul Yoo, "School Discipline and Racial Disparities in Early Adulthood," 6 2021. CES Working Paper Series.

Dee, Thomas S, "Are there civic returns to education?," *Journal of public economics*, 2004, *88* (9-10), 1697–1720.

_ , "A teacher like me: Does race, ethnicity, or gender matter?," *American Economic Review*, 2005, *95* (2), 158–165.

Girvan, Erik J, Cody Gion, Kent McIntosh, and Keith Smolkowski, "The relative contribution of subjective office referrals to racial disproportionality in school discipline.," *School Psychology Quarterly*, 2017, *32* (3), 392.

Holt, Stephen B and Seth Gershenson, "The impact of demographic representation on absences and suspensions," *Policy Studies Journal*, 2019, *47* (4), 1069–1099.

Jr, Roland G Fryer, "An empirical analysis of racial differences in police use of force," *Journal of Political Economy*, 2019, *127* (3), 1210–1261.

Kinsler, Josh, "Understanding the black–white school discipline gap," *Economics of Education Review*, 2011, *30* (6), 1370–1383.

Lindsay, Constance A and Cassandra MD Hart, "Exposure to same-race teachers and student disciplinary outcomes for Black students in North Carolina," *Educational Evaluation and Policy Analysis*, 2017, *39* (3), 485–510.

Lochner, Lance and Enrico Moretti, "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports," *American economic review*, 2004, *94* (1), 155–189.

Mendez, Linda M Raffaele and Howard M Knoff, "Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district," *Education and Treatment of Children*, 2003, pp. 30–51.

Miller, Douglas L, Na'ama Shenhav, and Michel Z Grosz, "Selection into identification in fixed effects models, with application to Head Start," Technical Report, National Bureau of Economic Research 2019.

Moretti, Enrico, "Workers' education, spillovers, and productivity: evidence from plant-level production functions," *American Economic Review*, 2004, *94* (3), 656–690.

Okonofua, Jason A, David Paunesku, and Gregory M Walton, "Brief intervention to encourage empathic discipline cuts suspension rates in half among adolescents," *Proceedings of the National Academy of Sciences*, 2016, *113* (19), 5221–5226.

Shi, Ying and Maria Zhu, "Equal Time for Equal Crime? Racial Bias in School Discipline," 2021.

Sorensen, Lucy C, Shawn D Bushway, and Elizabeth J Gifford, "Getting tough? the effects of discretionary principal discipline on student outcomes," *Education Finance and Policy*, Forthcoming.

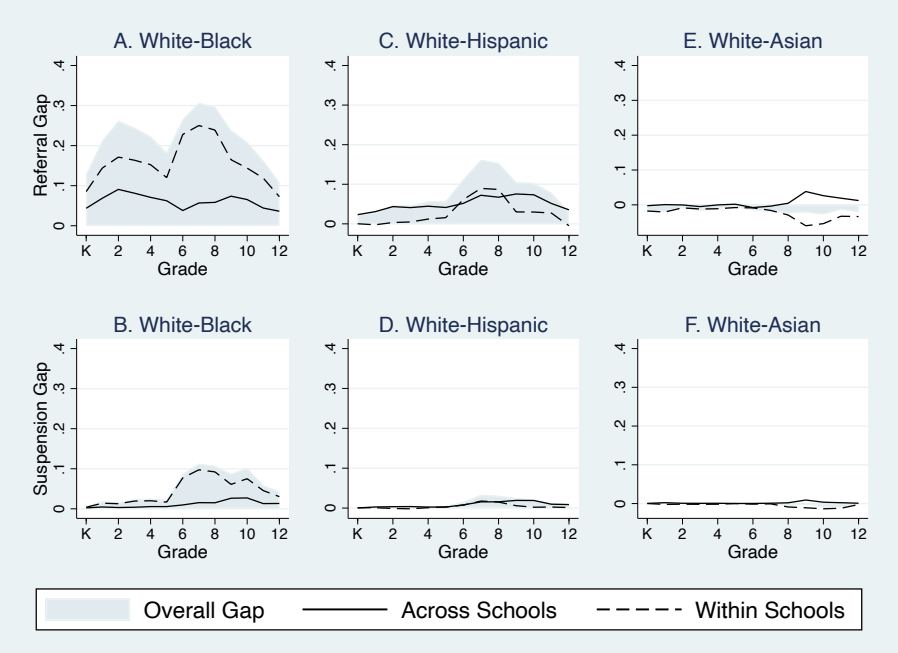
Steinberg, Matthew P and Johanna Lacoe, "What do we know about school discipline reform? Assessing the alternatives to suspensions and expulsions," *Education Next*, 2017, *17* (1), 44–53.

Sum, Andrew, Ishwar Khatiwada, Joseph McLaughlin, and Sheila Palma, "The consequences of dropping out of high school," *Center for Labor Market Studies Publications*, 2009, *23.*

Weisburst, Emily K, "Patrolling public schools: The impact of funding for school police on student discipline and long-term education outcomes," *Journal of Policy Analysis and Management*, 2019, *38* (2), 338–365.

Figures and Tables

Figure 1: Racial Gaps in the Likelihood of Receiving a Referral and Suspension this Year



Notes: This figure shows the decomposition of the raw racial gaps in referrals and suspensions by grade. Data come from a large urban school district in California from school years 2016-17 to 2019-2020. Technical details of the decomposition are documented in Appendix B.

Table 1: Student-by-Year Descriptive Statistics

	All	Race Comparison				
	Students	White	Black	Hispanic	Asian	Other
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Student characteristics						
White	0.12	1.00				
Black	0.07		1.00			
Hispanic	0.30			1.00		
Asian	0.33				1.00	
Other Race	0.18					1.00
Female	0.48	0.49	0.49	0.47	0.48	0.48
Special Education	0.16	0.14	0.29	0.20	0.10	0.14
Elementary School	0.41	0.45	0.36	0.42	0.34	0.54
Middle School	0.21	0.21	0.24	0.21	0.24	0.14
High School	0.30	0.24	0.34	0.29	0.37	0.18
Missing Grade-Level	0.08	0.09	0.05	0.07	0.05	0.13
Resides in Poorest Neighborhood	0.22	0.07	0.51	0.27	0.18	0.20
Resides in Poor Neighborhood	0.24	0.20	0.19	0.27	0.25	0.21
Resides in Less Poor Neighborhood	0.22	0.25	0.12	0.19	0.25	0.22
Resides in Least Poor Neighborhood	0.24	0.40	0.11	0.18	0.24	0.27
Missing Poverty Data	0.09	0.09	0.08	0.09	0.07	0.10
Lagged Non-cumulative GPA	3.13	3.23	2.86	2.96	3.28	3.15
	[0.54]	[0.41]	[0.70]	[0.60]	[0.50]	[0.39]
Missing Lagged GPA data	0.64	0.69	0.64	0.67	0.50	0.78
Panel B: Disciplinary outcomes						
At least one referral	0.08	0.05	0.26	0.12	0.03	0.07
Total referrals	0.37	0.13	1.77	0.49	0.08	0.33
	[2.67]	[1.16]	[6.33]	[2.78]	[0.84]	[2.76]
Total referrals conditional on	4.56	2.93	6.91	4.08	2.48	4.92
at least one referral	[8.24]	[4.93]	[10.98]	[7.08]	[4.00]	[9.57]
At least one suspension	0.02	0.01	0.07	0.02	0.01	0.01
Total suspensions	0.02	0.01	0.11	0.03	0.01	0.02
	[0.25]	[0.13]	[0.58]	[0.14]	[0.12]	[0.21]
Total suspensions conditional	1.58	1.34	1.76	1.55	1.36	1.54
on at least one suspension	[1.36]	[0.89]	[1.50]	[1.37]	[0.98]	[1.40]
Total suspended days conditional	3.22	2.69	3.71	3.07	2.74	3.14
on at least one suspension	[3.18]	[2.24]	[3.49]	[3.06]	[2.86]	[3.24]
Ratio of Suspensions to Referrals	0.05	0.04	0.07	0.05	0.05	0.04
	[0.18]	[0.16]	[0.19]	[0.17]	[0.19]	[0.15]
Total Observations	240,652	29,142	17,232	71,044	79,262	43,972

Notes: Standard deviations are reported in brackets for all non-binary variables. Data come from a large urban school district in California between the 2016-17 to 2019-20 school years. The unit of analysis is at the student-by-year level. There are 240,652 student-by-year observations. The "other" race category includes multiracial students and student missing race data. All the statistics above are reported as proportions, except for the lagged GPA scores, the total referrals, total suspensions, total suspended days, and ratio of suspensions to referrals.

Table 2: Racial Gaps in Annual Student Discipline Outcomes

	At least one ...		Total Number of ...		Conversion
	Suspension	Referral	Suspensions	Referrals	Rate
	(1)	(2)	(3)	(4)	(5)
Black	0.032***	0.101***	0.143***	2.022***	0.018***
	(0.004)	(0.009)	(0.024)	(0.270)	(0.005)
Hispanic	-0.001	0.013***	-0.015	0.147	-0.004
	(0.001)	(0.004)	(0.018)	(0.165)	(0.004)
Other Race	0.002	0.005*	0.053*	0.560**	0.003
	(0.001)	(0.003)	(0.030)	(0.222)	(0.006)
Asian	-0.004***	-0.021***	-0.022	-0.354**	-0.001
	(0.001)	(0.004)	(0.018)	(0.150)	(0.007)
Missing Race	0.002*	0.011***	0.035*	1.048***	0.001
	(0.001)	(0.004)	(0.021)	(0.279)	(0.005)
White Student Mean	0.007	0.046	0.130	2.934	0.038
Controls for:					
School-Year FEs	✓	✓	✓	✓	✓
Time-varying controls	✓	✓	✓	✓	✓
Adjusted R-squared	0.070	0.158	0.095	0.143	0.010
Observations	240,652	240,652	19,697	19,697	19,697

Notes: Clustered-robust standard errors at the school level are in parentheses. The omitted race group is white students. The conversion rate is the ratio of total suspensions to total referrals. The time-varying controls include gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. Columns 3 through 5 include only students with at least one referral. p<0.10* p<0.05** p<0.01***.

Table 3: Within-Incident Racial Disparities in Disciplinary Outcomes: Suspension

	(1)	(2)	(3)
	Panel A – Race Categories Consolidated		
Black/Hispanic/Other	0.022**	0.021**	0.018**
	(0.010)	(0.010)	(0.008)
Asian	0.010	0.012	0.009
	(0.011)	(0.011)	(0.010)
	Panel B – Detailed Race Categories		
Black	0.028***	0.027***	0.023***
	(0.010)	(0.010)	(0.009)
Hispanic	0.016	0.016	0.014
	(0.010)	(0.010)	(0.009)
Other	0.025**	0.024**	0.019**
	(0.011)	(0.011)	(0.010)
Asian	0.011	0.012	0.009
	(0.011)	(0.011)	(0.010)
White Student Mean		0.026	
Controls:			
Incident FEs	✓	✓	✓
Student Characteristics		✓	✓
Prior Student Achievement		✓	✓
Prior Year's Discipline		✓	✓
Current Year's Discipline			✓
Unique Multi-Race Referrals	12,277	12,277	12,277
Unique Multi-Race Incidents	5,195	5,195	5,195
Unique All Referrals	20,519	20,519	20,519
Unique All Incidents	9,012	9,012	9,012

Notes: Clustered-robust standard errors at the school level are in parentheses. Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The omitted group is white students. The "minoritized" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial, American Indian, Arabic, and Samoan students. The student characteristics includes gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. All model specifications include a race category called "missing race" for those students missing race data. $p<0.10$* $p<0.05$** $p<0.01$***.

Table 4: Within-Incident Racial Disparities in Disciplinary Outcomes by Incident Type

	All	Violence	Drugs	Interper	Defiance	Walkout
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A – Race Categories Consolidated					
Black/Hispanic/Other	0.018**	0.031**	0.042	0.018	0.001	-0.006
	(0.008)	(0.013)	(0.073)	(0.023)	(0.018)	(0.007)
Asian	0.009	0.035**	-0.069	0.008	-0.010	-0.036*
	(0.010)	(0.016)	(0.091)	(0.026)	(0.018)	(0.021)
	Panel B – Detailed Race Categories					
Black	0.023***	0.034**	0.015	0.028	0.004	0.004
	(0.009)	(0.014)	(0.086)	(0.024)	(0.018)	(0.010)
Hispanic	0.014	0.030**	0.065	0.009	-0.004	-0.011
	(0.009)	(0.014)	(0.071)	(0.023)	(0.018)	(0.009)
Other	0.019**	0.030**	-0.031	0.014	0.011	-0.012
	(0.009)	(0.014)	(0.114)	(0.030)	(0.020)	(0.012)
Asian	0.009	0.036**	-0.046	0.007	-0.010	-0.036*
	(0.010)	(0.016)	(0.092)	(0.026)	(0.018)	(0.021)
White Student Mean	0.026	0.027	0.080	0.021	0.034	0.000
Multi-Race Referrals	12,277	3,570	109	2,351	4,118	1,989
Multi-Race Incidents	5,195	1,972	56	1,498	2,133	1,019
Referrals	20,519	5,877	238	3,939	6,889	3,335
Incidents	9,012	3,250	116	2,515	3,689	1,755

Notes: Clustered-robust standard errors at the school level are in parentheses. Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The omitted group is white students. The "minoritized" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial, American Indian, Arabic, and Samoan students students. All models include incident fixed effects, student characteristics, prior student achievement, and prior student discipline. The student characteristics includes gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. All model specifications include a race category called "missing race" for those students missing race data. $p<0.10$* $p<0.05$** $p<0.01$***.

Appendix A

Table A1: Frequency of Referrals by Race and School Level

	White	Black	Hispanic	Asian	Other
Elementary	1,328	9,948	9,290	1,414	6,886
	1.56%	11.72%	10.94%	1.67%	8.11%
Middle	1,376	12,781	15,838	2,853	3,896
	1.62%	15.05%	18.65%	3.36%	4.59%
High School	905	6,397	8,034	1,884	2,077
	1.07%	7.53%	9.46%	2.22%	2.45%
All	3,609	29,126	33,162	6,151	12,859
	4.25%	34.30%	39.06%	7.24%	15.14%

Note: The unit of analysis is at the referral level. The other race category includes both multi-race students and students missing race data.

Table A2: Frequency of Referrals by Reason and School Level

	Violence	Drugs	Interpersonal Offenses	Disruption/ Noncompliance	Class Skipping or Walkout	Other Reason	Total
Elem	14,696	29	5,781	6,747	1,342	271	28,866
	17.31%	0.03%	6.81%	7.95%	1.58%	0.32%	34.00%
Middle	7,693	181	10,289	12,527	5,590	464	36,744
	9.06%	0.21%	12.12%	14.75%	6.58%	0.55%	43.28%
High	1,846	667	5,312	7,511	3,740	221	19,297
	2.17%	0.79%	6.26%	8.85%	4.40%	0.26%	22.73%
All	24,235	877	21,382	26,382	10,672	956	84,907
	28.54%	1.03%	25.18%	31.55%	12.57%	1.13%	100.00%

Note: The unit of analysis is at the referral level.

Table A3: Comparing Characteristics of Different Types of Incidents

	Type of Incidents				
Variables	(1) Single Student	(2) Multi-Student, Same-Race	(3) Multi-Student, Multi-Race	(4) P-value (1)=(3)	(5) P-value (2)=(3)
# of Students	1.00	2.27	2.75	0.00	0.00
White	0.05	0.02	0.05	0.34	0.00
Black	0.34	0.39	0.31	0.00	0.00
Hispanic	0.38	0.52	0.32	0.00	0.00
Asian	0.07	0.03	0.10	0.00	0.00
Other race	0.08	0.01	0.11	0.00	0.00
Female	0.25	0.32	0.32	0.00	0.69
Special Education	0.43	0.30	0.32	0.00	0.00
Lagged GPA	2.60	2.46	2.62	0.03	0.00
Missing Lagged GPA data	0.55	0.43	0.51	0.00	0.00
Lagged # of Referrals	7.08	5.80	5.76	0.00	0.75
Lagged # of Suspensions	0.37	0.30	0.26	0.00	0.08
Elementary School	0.34	0.22	0.29	0.00	0.00
Middle School	0.38	0.53	0.48	0.00	0.00
High School	0.21	0.23	0.21	0.02	0.00
Violence	0.31	0.28	0.29	0.00	0.17
Drugs	0.01	0.02	0.01	0.70	0.00
Interpersonal Offenses	0.26	0.20	0.19	0.00	0.46
Disruption/Noncompliance	0.30	0.33	0.34	0.00	0.88
Class Skipping/Walkout	0.11	0.16	0.16	0.00	0.87
Other Reasons	0.01	0.01	0.01	0.41	0.91
# of Observations	68,565	8,795	12,273		
# of Unique Incidents	68,565	4,370	5,192		

Notes: This table compares characteristics of three types of incidents in our sample. Columns 4 and 5 provide p values for simple two-sample T tests comparing single-student and multi-student, same-race incidents to our identifying sample which is multi-student multi-race incidents.

Table A4: Racial Gaps in Annual Student Discipline Outcomes (Simple Model)

	At least one ...		Total Number of ...		Conversion
	Suspension	Referral	Suspensions	Referrals	Rate
	(1)	(2)	(3)	(4)	(5)
Black	0.059***	0.211***	0.248***	3.978***	0.030***
	(0.007)	(0.023)	(0.036)	(0.580)	(0.007)
Hispanic	0.013***	0.073***	0.067***	1.148***	0.009
	(0.002)	(0.012)	(0.025)	(0.257)	(0.006)
Other Race	0.005***	0.023***	0.084***	1.585***	0.005
	(0.002)	(0.005)	(0.032)	(0.353)	(0.006)
Asian	-0.001	-0.014***	0.014	-0.457**	0.011
	(0.001)	(0.005)	(0.023)	(0.186)	(0.008)
Missing Race	0.002*	0.019***	0.025	2.403***	-0.006
	(0.001)	(0.006)	(0.022)	(0.401)	(0.006)
White Student Mean	0.007	0.046	0.130	2.934	0.038
Adjusted R-squared	0.016	0.048	0.011	0.031	0.004
Observations	240,652	240,652	19,697	19,697	19,697

Notes: Clustered-robust standard errors at the school level are in parentheses. The omitted race group is white students. The conversion rate is the ratio of total suspensions to total referrals. None of the regressions above include fixed effects, or control variables. Columns 3 through 5 include only students with at least one referral. p<0.10* p<0.05** p<0.01***.

Table A5: Within-Incident Racial Disparities in Disciplinary Outcomes: Suspension Days

	(1)	(2)	(3)
	Panel A – Race Categories Consolidated		
Minoritized	0.049*	0.045*	0.039*
	(0.025)	(0.025)	(0.023)
Asian	0.016	0.020	0.014
	(0.029)	(0.029)	(0.027)
	Panel B – Detailed Race Categories		
Black	0.059**	0.055**	0.046*
	(0.026)	(0.026)	(0.024)
Hispanic	0.042*	0.039	0.035
	(0.025)	(0.025)	(0.023)
Other	0.046*	0.044	0.034
	(0.027)	(0.027)	(0.025)
Asian	0.016	0.020	0.015
	(0.029)	(0.029)	(0.027)
White Student Mean		0.038	
Controls:			
Incident FEs	✓	✓	✓
Student Characteristics		✓	✓
Prior Student Achievement		✓	✓
Prior Year's Discipline		✓	✓
Current Year's Discipline			✓
Unique Multi-Race Referrals	12,277	12,277	12,277
Unique Multi-Race Incidents	5,195	5,195	5,195
Unique All Referrals	20,519	20,519	20,519
Unique All Incidents	9,012	9,012	9,012

Notes: Clustered-robust standard errors at the school level are in parentheses. Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The omitted group is white students. The "minoritized" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial, American Indian, Arabic, and Samoan students. The student characteristics includes gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. All model specifications include a race category called "missing race" for those students missing race data. $p<0.10$* $p<0.05$** $p<0.01$***.

Table A6: Regressions on Likelihood of Suspension and Suspension Days

	Likelihood of Suspension			Suspension Days		
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A – Race Categories Consolidated					
Minoritized	0.017	0.017	0.017	0.046	0.044	0.044
	(0.012)	(0.012)	(0.012)	(0.032)	(0.028)	(0.028)
Asian	0.017	0.018	0.018	0.029	0.029	0.029
	(0.015)	(0.015)	(0.015)	(0.034)	(0.032)	(0.032)
White Student Mean		0.012			0.010	
Controls:						
Incident FEs	✓	✓	✓	✓	✓	✓
Student Characteristics		✓	✓		✓	✓
Prior Student Achievement		✓	✓		✓	✓
Prior Year's Discipline		✓	✓		✓	✓
Current Year's Discipline			✓			✓
Unique Multi-Race Referrals	3,169	3,169	3,169	3,169	3,169	3,169
Unique Multi-Race Incidents	2,115	2,115	2,115	2,115	2,115	2,115
Unique All Referrals	5,396	5,396	5,396	5,396	5,396	5,396
Unique All Incidents	3,653	3,653	3,653	3,653	3,653	3,653

Notes: Cluster-robust standard errors at the incident level are in parentheses. Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The sample only includes observations where the student has no prior referrals this school year. The omitted group is white students. The "minority" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial students. The student characteristics include gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. All model specifications include a race category called "missing race" for those students missing race data. $p<0.10^{*}$ $p<0.05^{**}$ $p<0.01^{***}$.

Table A7: Weighted Regression Results Accounting for Selection into Identification

	Likelihood of Suspension		Suspension Days	
	(1)	(2)	(3)	(4)
	Unweighted	Weighted	Unweighted	Weighted
Minoritized	0.007*	0.008*	0.014+	0.017+
	(0.004)	(0.004)	(0.009)	(0.009)
R2	0.643	0.524	0.622	0.509
Observations	20,512	6,057	20,512	6,057

Notes: Different from our main specification, we combine Asian and white students as the referrence group so we only have one treatment group in order to implement the weighting strategy. Following Miller et al. (2019), we implement a one-step weighting strategy that uses the product of predicted likelihood of being in the identifying sample and inverse conditional variance as regression weights. $p<0.10$* $p<0.05$** $p<0.01$***.

Table A8: Regressions on Likelihood of Suspension by School Level

	All	Elem	Middle	High
	(1)	(2)	(3)	(4)
	Panel A – Race Categories Consolidated			
Minority	0.018**	0.007	0.008	0.043**
	(0.008)	(0.010)	(0.016)	(0.017)
Asian	0.009	0.020*	-0.005	0.027
	(0.010)	(0.011)	(0.018)	(0.018)
	Panel B – Detailed Race Categories			
Black	0.023***	0.005	0.015	0.054***
	(0.009)	(0.010)	(0.016)	(0.018)
Hispanic	0.014	0.004	0.002	0.040**
	(0.009)	(0.011)	(0.016)	(0.017)
Other	0.019**	0.016	0.014	0.025
	(0.009)	(0.010)	(0.018)	(0.021)
Asian	0.009	0.020*	-0.005	0.028
	(0.010)	(0.011)	(0.018)	(0.018)
White Student Mean	0.026	0.010	0.047	0.016
Controls:				
Incident FEs	✓	✓	✓	✓
Student Characteristics	✓	✓	✓	✓
Prior Student Achievement	✓	✓	✓	✓
Prior Discipline	✓	✓	✓	✓
Unique Multi-Race Referrals	12,277	3,534	5,933	2,522
Unique Multi-Race Incidents	5,195	1,597	2,413	1,084
Unique Referrals	20,519	5,356	10,338	4,400
Unique Incidents	9,012	2,475	4,431	1,954

Notes: Clustered-robust standard errors at the school level are in parentheses. Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The omitted group is white students. The "minority" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial students. The student characteristics includes gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. All model specifications include a race category called "missing race" for those students missing race data. p<0.10* p<0.05** p<0.01***.

Table A9: Heterogeneity Results

	All	High School	All	High School	All	High School
	(1)	(2)	(3)	(4)	(5)	(6)
Minority	0.030	0.044*	0.015	0.042**	0.025**	0.044**
	(0.019)	(0.026)	(0.009)	(0.017)	(0.011)	(0.022)
Asian	0.038*	0.044	0.006	0.024	0.008	0.015
	(0.021)	(0.030)	(0.010)	(0.019)	(0.012)	(0.023)
Male	0.017	0.009				
	(0.020)	(0.034)				
Minority × Male	-0.017	-0.003				
	(0.020)	(0.034)				
Asian × Male	-0.039*	-0.025				
	(0.023)	(0.037)				
Special Education			-0.007	-0.007		
			(0.019)	(0.055)		
Minority × Spec-Ed			0.011	-0.001		
			(0.019)	(0.055)		
Asian × Spec-Ed			0.010	0.013		
			(0.023)	(0.054)		
Poor					0.011	-0.005
					(0.015)	(0.030)
Minority × Poor					-0.017	-0.003
					(0.015)	(0.030)
Asian × Poor					-0.005	0.016
					(0.016)	(0.032)
White Student Mean	0.026	0.016	0.026	0.016	0.026	0.016
Unique Multi-Race Referrals	12,277	2,522	12,277	2,522	12,277	2,522
Unique Multi-Race Incidents	5,195	1,084	5,195	1,084	5,195	1,084
Unique Referrals	20,519	4,400	20,519	4,400	20,519	4,400
Unique Incidents	9,012	1,954	9,012	1,954	9,012	1,954

Notes: Data come from a large urban school district in California from school year 2016-17 to 2019-20. The unit of analysis is at the incident level. The omitted group is white students. The "minority" category includes black, Hispanic, and "other" race students. The "other" race category includes multiracial students. All models include incident fixed effects (FEs), student characteristics, prior student achievement, and prior student discipline. The student characteristics includes gender, special education status, grade-level, student's neighborhood poverty-rate, lagged non-cumulative GPA, and lagged student discipline outcomes. The "poor" category includes students residing in neighborhoods that have poverty rates below the 50th percentile. All model specifications include a race category called "missing race" for those students missing race data.

Appendix B: Decompositions

Decomposing Racial Gaps

We decompose racial gaps in referrals and suspensions into between-school and within-school components. We compare Black, Hispanic, and Asian students to their white peers by using both the likelihood of receiving a referral and the likelihood of having a suspension in a school year as our two outcomes. Following Barret et al. (2019), we define the raw average referral or suspension rate $\bar{D}_{is}$ for a given group of students in a given grade weighted across students and schools using equation (1) below:

$$\bar{D}_{is} = \frac{\sum_i \sum_s Group_{is} Y_{is}}{\sum_i \sum_s Group_{is}} \tag{1}$$

where i indicates students and s indicates schools. $Group_{is}$ indicates the student's racial or ethnic identity. Y_{is} takes the value of 1 if the student receives, for example, an office referral in the focal year, and 0 otherwise.

For simplicity, we use $\bar{D}_{is}$ to represent white students' referral or suspension rates and $\tilde{D}_{is}$ is to indicate the same measure for a non-white student group, which can be Black, Hispanic, or Asian students. Our goal is to decompose the raw gap $\bar{D}_{is}$-$\tilde{D}_{is}$ into between- and within-school components using equation (2) below:

$$\bar{D}_{is} - \tilde{D}_{is} = \bar{D}_s - \tilde{D}_s + ((\bar{D}_{is} - \tilde{D}_{is}) - (\bar{D}_s - \tilde{D}_s)) \tag{2}$$

$\bar{D}_s - \tilde{D}_s$ would be the measure on between-school gap and $(\bar{D}_{is} - \tilde{D}_{is}) - (\bar{D}_s - \tilde{D}_s)$ is the within-school gap. To plot Figure 1, we compute elements in Equation (2) for each grade (K-12) and each minoritized-white combinations for both referral and suspension rates using our analytic sample (school years 2016-2017 to 2019-2020).

Appendix C: Adapting Miller et al. (2019) Weights

Re-weighted FE Estimator Considering Selection into Identification

Our goal is to identify the "causal effect" of race on students' suspension outcomes. The selection into identification issue arises when treatment status only varies in certain incidents, which induces a non-random selection of incidents into the identifying sample and causes bias. Specifically, this means that being involved in a multi-student multi-race incident might be correlated with some fixed characteristics of an incident. For example, if Black students are more likely to be involved in fight incidents with white students, our incident FE model would mainly draw on variation from fight incidents. As Black students are over-represented in fight incidents, our estimate would be upward biased.

Following the terms used by Miller et al. (2019), let $D_i \in \{0, 1\}$ indicate whether a student i is involved in a multi-student multi-race incident (i.e., our "treatment") and $g(i)$ be the relevant group (i.e., incident) for i. Our target population T is all students. We denote incidents that have variation on student race/ethnicity (i.e., $(\text{Var}\left(D_i \mid i \in g(i)\right) > 0)$) as "switchers" S .

We can use two propensity scores constructed from the vector of incident characteristics, $\mathbf{X_g}$, as the conditioning variables to define the likelihood of an individual student to be in a switching group (S_g) or a target group (Q_g): $P_x := \Pr\left[S_g = 1 \mid \mathbf{X_g} = \mathbf{x}\right]$ and $Q_x := [T_g = 1 \mid \mathbf{X_g} = \mathbf{x}]$.

The re-weighted FE estimator for our target population t can be specified as

$$\widehat{\delta^t} := \frac{1}{\sum_i \mathbf{1}\left(S_{g(i)} = 1\right)} \sum_{i|S_{g(i)}=1} \widehat{w^t_{g(i)}} \cdot \hat{\delta}_{g,FE}$$

with $\widehat{w^t_{g(i)}}$ our estimate of $w^t_{g(i)}$,

$$w^t_{g(i)} := \frac{Q_x \cdot \Pr\left[S_g = 1\right]}{P_x \cdot \Pr\left[T_g = 1\right]}$$

Under a few assumptions (for details, see (Miller et al., 2019)), $\widehat{\delta^t}$ is unbiased for the average treatment effect of the target population. Intuitively, we upweight observations that are more similar to the target, and downweight observations that are overrepresented in the switching population.