Economics of Education Review 102 (2024) 102578

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Does the salience of race mitigate gaps in disciplinary outcomes? Evidence from school fights

Kyle Raze [a,1], Glen R. Waddell [b,c,*]

[a] U.S. Census Bureau, United States of America
[b] University of Oregon, United States of America
[c] IZA Bonn, Germany

ARTICLE INFO

JEL classification:
I24
J71

Keywords:
Student behavior
School discipline
Racial disparities

ABSTRACT

Racial gaps in the adjudication of student misconduct are well documented—relative to white students engaged in similar behaviors, students of color are more likely to be disciplined and the discipline they receive tends to be harsher. We show that racial disparities in the adjudication of fighting infractions depend on the racial composition of incidents. While significant disparities exist within schools, we find little if any within-incident disparities. Examining disparities across fights, we show that students of color are punished more severely, on average, as fights involving only students of color are punished more severely than fights involving only white students. Moreover, students of color in multi-race fights receive punishments that are statistically indistinguishable from those assigned to white students in fights involving only white students, suggesting that disparities arise from the differential adjudication of incidents by their racial composition rather than from the differential adjudication of students within the same incident.

1. Introduction

Childhood environments shape racial disparities in a variety of social and economic outcomes (Chetty, Dobbie, Goldman, Porter, & Yang, 2024; Chetty, Hendren, Jones, & Porter, 2020). With potential gains to correcting early differences in the experiences of those of different racial backgrounds, school environments are an important setting to consider—it is in these formative years that students are making decisions about investments in human capital and forming expectations of their own comparative advantages.[2] In this paper, we examine patterns of racial disparities in the adjudication of student misconduct and provide evidence of a mechanism that potentially explains their origins.

While school discipline can mitigate externalities associated with disruptive behavior (Carrell & Hoekstra, 2010; Kinsler, 2013; Pope & Zuo, 2023), disciplinary interventions impose significant costs on disciplined students. For example, the disciplinary actions commonly available to school administrators are inseparable from interruptions to the direct inputs into the production of human capital; suspensions, expulsions, and other forms of exclusionary discipline decrease instructional time and disrupt the continuity of instruction. As a result, exclusionary discipline can hinder academic performance (Anderson, Ritter, & Zamarro, 2019; Bacher-Hicks, Billings, & Deming, 2019; Craig & Martin, 2023; Sorensen, Bushway, & Gifford, 2022; Steinberg & Lacoe, 2018). Exposure to harsh exclusionary discipline regimes has also been shown to decrease educational attainment and increase the likelihood of arrest and incarceration (Bacher-Hicks et al., 2019). Equity concerns notwithstanding, racially biased discipline could therefore lead to significant and long-lasting economic inefficiencies in the production of human capital.

The existence of race-based disparities in disciplinary outcomes is well documented (Anderson & Ritter, 2017; Barrett, McEachin, Mills, & Valant, 2021; Gopalan & Nelson, 2019; Kinsler, 2011; Liu, Hayes, & Gershenson, 2024; Ritter & Anderson, 2018; Shi & Zhu, 2022; Welsh & Little, 2018). Existing research suggests that a significant portion of the average discipline gap between white students and students of color arises across schools, as Black students are more likely to live in school districts with higher rates of exclusionary discipline (Anderson & Ritter, 2017; Barrett et al., 2021; Gopalan & Nelson, 2019; Kinsler, 2011; Ritter & Anderson, 2018). Yet, gaps typically remain after conditioning on student characteristics and school fixed effects (Anderson & Ritter, 2020; Barrett et al., 2021; Beck & Muschkin, 2012; Gopalan & Nelson,

* Corresponding author at: University of Oregon, United States of America.
E-mail addresses: kyle.raze@census.gov (K. Raze), waddell@uoregon.edu (G.R. Waddell).

[1] Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau.

[2] For example, individual expectations about comparative advantage have been shown to influence consequential decisions about major choice in university settings (Arcidiacono, Aucejo, & Spenner, 2012; Card & Payne, 2021).

https://doi.org/10.1016/j.econedurev.2024.102578
Received 19 September 2023; Received in revised form 2 August 2024; Accepted 8 August 2024

2019; Liu et al., 2024; Shi & Zhu, 2022), which leaves open the possibility that school officials treat students of color more harshly than white students who engage in similar behaviors. Even so, within-school comparisons need not isolate the response of school officials to the race of their students—within-school disparities are consistent with differential treatment on the basis of race, but also with systematic but unobserved differences in student behavior.

To better adjust for unobserved differences in student behavior, recent studies compare the punishments of students implicated together in the same incident. Barrett et al. (2021) uses administrative data from Louisiana to compare the suspension lengths associated with incidents in which exactly two students were suspended for fighting on the same day in the same school. Shi and Zhu (2022) leverages the availability of incident identifiers in North Carolina to make within-incident comparisons. (As will be the case in our setting, the North Carolina data include cases that did not end in punishment, and having an incident identifier circumvents the need for a same-day, same-school, same-incident-type matching rule.) Likewise, Liu et al. (2024) leverages incident identifiers in administrative data on student referrals, but from a large California school district. All three studies identify small but statistically significant within-incident differences in the number of days suspended (i.e., among fights that end in suspension in the case of Barrett et al., 2021, and among all incidents in Liu et al., 2024; Shi & Zhu, 2022). Together, these studies suggest that school administrators impose harsher punishments on students of color when adjudicating cases of student misconduct. In terms of magnitude, they each find similarly-sized racial disparities in suspension lengths—Black students are suspended for roughly one twentieth of a day longer than white students implicated in the same incident.[3]

That being said, within-incident race differentials are identified off of a very specific subset of incidents—those in which there was variation in race. Thus, any identification strategy that relies on incident fixed effects necessarily decouples those incidents that have variation in race from those that do not. Moreover, it is when administrators adjudicate students of color and white students "side-by-side" that one would expect racial differences to be more salient, which may induce disciplinary outcomes that are not representative of the adjudication of students of color in other contexts. For example, variation in the racial composition of incidents may lead to "signal jamming" behavior (Fudenberg & Tirole, 1986; Holmström, 1999), by which administrators anticipate that the most reliable signal of their treatment of race is likely to be found in their adjudication of multi-race incidents. If there are professional repercussions to exhibiting explicit biases, it would be in administrators' interest to pay closer attention to their treatment of students of color, which could move them toward equal treatment in multi-race incidents in particular. Or perhaps administrators more easily suppress implicit biases when differences in race are more evident, which again implies that racial disparities in adjudication outcomes would vary across the racial composition of incidents. The psychology literature on preference reversals in joint evaluation (Bazerman, Moore, Tenbrunsel, Wade-Benzoni, & Blount, 1999; Hsee, Loewenstein, Blount, & Bazerman, 1999) also suggests that racial composition may directly matter to outcomes, insofar as the adjudication of students of color apart from white students leads decision makers to put less weight on equal treatment (i.e., a "difficult-to-evaluate attribute" in the spirit of Hsee et al., 1999) than they would in the joint evaluation of students of color and white students side-by-side.

From a variety of perspectives, there is good reason to anticipate direct effects of racial composition on outcomes. One of the takeaways from our analysis will be that existing gaps in adjudication outcomes are not seeming to arise from differential treatment within multi-race incidents. While potential within-incident differences are important and informative, this narrow source of variation overlooks the consequences of other relevant decisions made by administrators and the broader experiences of students of color in the adjudication process.

We therefore seek to contextualize the adjudication of multi-race incidents by comparing adjudication outcomes across joint incidents that are similar but still vary in racial composition. Using data from Washington, we exploit the availability of incident identifiers and a well-defined infraction category ("fighting without major injury") to identify how school administrators respond to the racial composition of joint incidents when adjudicating student misconduct. In multi-race fights, we find no race-based differences in outcomes. However, when we inquire into the attenuation of within-school disparities upon controlling for incident fixed effects, we identify large within-school differences across same-race fights wherein fights involving only students of color elicit harsher punishments than those involving only white students. In high schools, for example, students of color in fights that involve *only* students of color receive suspensions that are two thirds of a day longer than those assigned to white students in all-white fights. Yet, being implicated with a white student renders the punishment of students of color indistinguishable from the punishment of white students in all-white fights. In other words, our results suggest that within-school disparities are driven by differences in the treatment of students of color *across* incidents, which implies that purging all within-incident disparities in punishment would do little to close the gap in disciplinary outcomes between students of color and their white peers.

While our point estimates of within-incident disparities tend to be smaller than the modest within-incident disparities documented in Barrett et al. (2021), Liu et al. (2024), and Shi and Zhu (2022), our main contribution to the literature is in demonstrating that school administrators adjudicate students of color much differently in incidents with variation in race than they do in incidents without variation in race. Within-incident comparisons foreclose on the opportunity to detect this pattern, as there is no variation in racial composition within incidents. The large disparities we observe across same-race fights suggest that differential treatment is greater across incidents than within incidents. If anything, the salience of differences in race within an incident moves administrators toward equal treatment.[4]

2. Data

To document racial differences in the severity of sanctions for alleged misconduct, we consider fighting infractions reported by public schools in Washington to the Office of the Superintendent of Public Instruction between 2014–15 and 2017–18.[5] Similar to Liu et al. (2024) and Shi and Zhu (2022), our administrative data include incident identifiers and infractions that did not result in students receiving

[3] Liu et al. (2024) and Shi and Zhu (2022) also document statistically significant differences for Black students on the extensive margin of suspension, with smaller effects in North Carolina (Shi & Zhu, 2022) than in California (Liu et al., 2024).

[4] We note that the results of the student fixed effects analysis in Shi and Zhu (2022) are not necessarily at odds with our findings. Shi and Zhu (2022) estimates a model that identifies how *within-incident* differences in the punishments assigned to students of color vary across the racial composition of incidents—Black students receive harsher punishments relative to the other student in the same incident when the other student is of a different racial background. However, by modeling outcomes as punishments *relative* to others in the same incident, rather than in levels, the student fixed effects analysis in Shi and Zhu (2022) implicitly absorbs level differences in punishment across incidents, leaving open the possibility that the severity of punishments (i.e., their levels) imposed on students of color decreases in multi-race incidents. Neither Barrett et al. (2021) nor Liu et al. (2024) estimate student fixed effects specifications.

[5] For our purposes, public schools include traditional public schools as well as public charters and alternative schools—we exclude infractions from special education schools and juvenile correctional institutions.

exclusionary discipline. We complement Barrett et al. (2021), Shi and Zhu (2022), and Liu et al. (2024) by identifying disparities in a setting in which racial resentment is persistently lower (Smith, Kreitzer, & Suo, 2020).[6] While there are still significant race-based gaps in the adjudication of outcomes in Washington, both across and within schools, one might reasonably expect school administrators in Washington to respond differently to race than administrators in other states.

While our data will facilitate the ability to identify fights, on other margins we will be limited. There is a degree of difficulty in measuring race categorically, and coarse racial categories prevent us from distinguishing between students who report more than one race. For example, while it is easy to imagine that students who identify as both Black and white experience different disciplinary outcomes than students who identify as both Asian and white, the data record both types as "two or more races". Similarly, the data do not allow us to distinguish between race and ethnicity, as students who report Hispanic ancestry are coded as "Hispanic", regardless of their race. As a result, the available racial categories can complicate the interpretation of specific racial gaps, as students perceived by administrators as one race (e.g., Black) may be coded in the data as another (e.g., Hispanic, or as two or more races). Moreover, the considerable racial diversity in the sample can limit our ability to precisely estimate specific gaps, such as those between monoracial Black and white students. Thus, to economize on statistical power, we conduct our main analysis around incidents that involve only white students, incidents that involve only students of color, and those that involve both white students and students of color, defining "students of color" as those who do not identify as white non-Hispanic.[7] That said, the qualitative conclusions from a more granular analysis of Black-white and Hispanic-white punishment disparities, which we provide in Section 5, are unchanged.

2.1. Sample selection

We restrict our attention to incidents that (i) are well defined, (ii) are more likely to have well-defined sets of participants, (iii) are narrow enough in scope that we can argue that any remaining racial disparities are not likely to be explained by incident heterogeneity, and (iv) are not so rare that they lack economic significance. A set of incidents that satisfies these criteria provides as close to as-good-as-random variation as possible while still allowing us to contextualize multi-race incidents with a set of similar, but same-race incidents.

To satisfy these criteria, we consider infractions for "fighting without major injury" among boys.[8] In addition to being included in mandatory federal reporting, the fights in our sample are well-defined by the state. State guidance defines "fighting without major injury" as "mutual participation in an incident involving physical violence" and specifically conditions on incidents in which no "persons on school grounds require professional medical attention" (Reykdal, Weaver-Randall, & Ireland, 2018). The state also provides examples of disqualifying injuries; fights that result in "stab or bullet wounds, concussions, fractured or broken bones, or cuts requiring stitches" would be adjudicated in a separate category of offense. Thus, if fights between students of color tend to be worse in some unobservable way that rationalizes harsher penalties, "worse" must not be so much worse as to imply "cuts requiring stitches". In that way, "worse" has an upper bound of "not so much worse that there are stitches". Moreover, the state directs school officials to exclude "verbal confrontations, tussles, or other minor confrontations". Collectively, these describe a fairly narrow band of student activity over which we can examine differences in adjudication outcomes between students of color and white students.

Relative to other forms of joint misconduct, it can also be argued that fights are the least likely to originate from race-based selection into the sample. Consider "disruptive conduct", for example, which the state defines as any behavior "that materially and substantially interferes with the educational process" (Reykdal et al., 2018). The relative subjectivity permitted in determining what constitutes disruptive conduct would leave much more room for race-based selection into infractions. In contrast, well-defined conditions and mandatory reporting supports that selection into fights is less likely to depend on the subjective judgments of teachers, so our focus on fights tips toward limiting potential measurement error in the classification of incidents. To the extent that there are concerns about selection into fights, those concerns should be heightened considerably in the analysis of other types of incidents. As for framing the external validity of comparisons across fights, we note that fights are the most common type of multi-student incident in our data, and while there will surely be some students who escape the eyes of teachers, the "jointness" of fights leaves us more confident that we have captured the set of relevant actors.[9]

There are a total of 66,331 fighting infractions among boys in our data. While schools are required to use the same incident identifier for incidents that involve multiple students, 34 percent of fighting infractions occur in schools that never report the same incident identifier for multiple students. Thus, our analysis will speak only to schools that follow the reporting guidelines.[10] In schools that do report matching incident identifiers, not all fighting infractions have an incident identifier that matches that of another student in the infraction data. As a worst case, one might imagine that white students systematically avoid fighting infractions, leaving an "excess" of students of color among the reported fights. If it is the less severe white infractions that select out of reporting, then the measurable within-incident race differentials would understate the extent of differential adjudication in our identifying sample. While this possibility is not unique to our setting, the safest inference going forward might be to interpret our estimated differentials as lower bounds of the effect of race on outcomes. In total, we observe 16,271 infractions from 7637 multi-student incidents that implicate at least two boys for fighting. To further ensure the comparability of the fights in our sample, we discard 576 infractions from multi-student incidents that include girls or that implicate other students for non-fighting behaviors, though our conclusions are not sensitive to these restrictions.

[6] This is also consistent with the data collected by Project Implicit, which suggests that implicit racial biases in Washington are the second lowest among US states. For more information, see Chris Mooney, "Across America, Whites Are Biased and They Don't Even Know It", *Washington Post*, 8 December 2014, https://www.washingtonpost.com/news/wonk/wp/2014/12/08/across-america-whites-are-biased-and-they-dont-even-know-it/ [Accessed 1 June 2022], and Jordan Axt, "Mapping Geographical Variation in Implicit Racial Attitudes", *Project Implicit*, https://implicit.harvard.edu/implicit/user/jaxt/blogposts/piblogpost005.html [Accessed 1 June 2022].

[7] Specifically, we define students of color as those who identify as (i) solely Black, (ii) Hispanic (of any race), (iii) solely Asian, (iv) solely Pacific Islander, (v) solely Native American, or (vi) two or more races.

[8] Relative to the boys in our data, girls are rarely implicated for fighting (boys' infractions for fighting outnumber those of girls by a ratio of four to one). Moreover, limiting our attention to boys allows us to sidestep concerns about selection into other-gender fights. For example, to incorporate gender-based selection into the analysis in Section 4, we would need to stratify the sample by gender and the gender composition of fights in addition to race and the racial composition of fights. However, given relative dearth of girls in our data, we lack the statistical power to estimate a model that fully interacts race and racial composition with gender and gender composition.

[9] Further, note that no state reports data on victims, to our knowledge, and to the extent the victim is observable to those adjudicating student conduct (but not to the econometrician), there may also be missing race components to the adjudication of other categories of misconduct. Considering fights between students—fights being well-defined and subject to mandatory reporting—mitigates such concerns.

[10] Schools that follow the reporting guidelines tend to be less white, more urban, and more economically disadvantaged (as measured by the fraction of students who qualify for free or reduced-price meals) than schools that do not.

Table 1
Summary statistics.

	Grades PK-5		Grades 6-8		Grades 9-12	
	SoC	White	SoC	White	SoC	White
Panel A: All fighting infractions						
Fraction receiving any exclusionary discipline	0.425	0.369	0.852	0.812	0.918	0.903
Total days suspended (mean)	0.615	0.509	2.033	1.684	3.463	2.905
Total days suspended (standard deviation)	1.063	0.977	1.861	1.588	2.588	2.244
Fraction receiving severe discipline	0.002	0.002	0.010	0.005	0.037	0.019
Observations	16,120	13,393	13,326	11,667	6676	5149
Incidents	14,814	12,925	11,658	10,645	5699	4675
Students	9295	7538	9330	8427	5375	4285
Schools	837	900	604	634	375	396
Panel B: All fighting infractions from schools that report at least one multi-student fight						
Fraction receiving any exclusionary discipline	0.413	0.368	0.869	0.841	0.923	0.908
Total days suspended (mean)	0.641	0.524	2.115	1.762	3.557	2.959
Total days suspended (standard deviation)	1.144	0.990	1.891	1.579	2.604	2.238
Fraction receiving severe discipline	0.002	0.003	0.011	0.004	0.040	0.019
Observations	7650	4367	8467	6127	4256	2764
Incidents	6351	3904	6800	5106	3284	2293
Students	4433	2617	5990	4567	3474	2363
Schools	331	320	324	324	195	192
Panel C: Multi-student fights						
Fraction receiving any exclusionary discipline	0.444	0.430	0.901	0.875	0.946	0.928
Total days suspended (mean)	0.608	0.537	2.065	1.714	3.597	2.923
Total days suspended (standard deviation)	0.880	0.802	1.708	1.407	2.478	2.048
Fraction receiving severe discipline	–	–	0.007	–	0.035	0.009
Observations	2986	1570	4366	3215	2154	1404
Incidents	1743	1113	2764	2222	1261	957
Students	2328	1307	3617	2732	1942	1303
Schools	296	279	289	288	176	178

Notes: Summary statistics of punishment outcomes considered in Sections 3 and 4 for students of color (SoC) and white students. Exclusionary discipline consists of either a suspension or an expulsion. The alternative to exclusionary discipline is either "no intervention" or "other intervention." Severe discipline consists of a suspension longer than 10 school days or an expulsion. The sample in Panel A consists of boys' infractions for "fighting without major injury," excluding those involving weapons, between 2014–15 and 2017–18. The sample in Panel B consists of boys' fighting infractions from schools that report fights with matching incident identifiers. The sample in Panel C consists of fighting infractions from multi-student fights that implicate at least two boys for fighting, but do not include girls or implicate other students for non-fighting behaviors. A fight is classified as "multi-student" if two or more students have a matching incident identifier.

2.2. Outcomes

We consider three margins of formal exclusionary discipline as outcomes for each infraction: (i) whether the student receives any exclusionary discipline (defined as a suspension or expulsion), (ii) the total number of school days the student is suspended, and (iii) whether the student receives severe discipline (defined as a suspension longer than 10 school days or an expulsion).[11] In Table 1 we provide average disciplinary outcomes by race and grade span for (i) all fighting infractions in Washington, (ii) all fighting infractions at schools that use the same incident identifier (across students) when multiple students are involved in individual fights, and (iii) all multi-student fights at these schools. While punishments vary significantly across grade spans, average punishment outcomes are more severe for students of color than for white students within each grade span and sample. As we will show in Section 3, most of these unconditioned differences are statistically significant. With the exception of exclusionary discipline rates in the elementary grades, the observed racial differences in outcomes within each grade span are similar across samples.

[11] The alternative to exclusionary discipline is either "no intervention" or "other intervention". While schools are not required to report infractions that do not result in suspension or expulsion, 95 percent of fighting infractions are from schools that report infractions (for fighting or other behaviors) that result in "no intervention" or "other intervention". Our measure of suspension length, "total days suspended", includes zero for infractions that did not result in the student being suspended from school. It also includes expulsions, which are assumed to be as long as the longest suspension. To limit the influence of exceptionally long suspensions, we Winsorize "total days suspended" at the 99th percentile of the unstratified sample of boys' fighting infractions (10 school days).

2.3. Student characteristics

We observe a total of 41,511 students in the full sample and 12,849 students in multi-student fights. Roughly 41 percent of students in multi-student fights are white (non-Hispanic), 27 percent are Latino (Hispanic origin of any race), 16 percent are Black, 9 percent report more than one race, 3 percent are Asian, 2 percent are Pacific Islander, and 2 percent are Native American. We document other characteristics of multi-student fights in Table 2.

We derive controls for three sets of student attributes—socioeconomic status, disability, and past achievement—from an extended panel that dates back to 2009–10.

(a) We control for socioeconomic status using persistent eligibility for free or reduced-price meals. Using up to nine years of data, we determine whether a student is (i) always eligible, (ii) sometimes eligible, or (iii) never eligible for free or reduced-price meals. In doing so, we follow others who have argued that persistent eligibility provides a better proxy for current household income than current eligibility (Michelmore & Dynarski, 2017). While there are significant racial disparities in socioeconomic status within each grade level, the vast majority of infractions from multi-student fights implicate students from low-income households—this is true for white students and students of color alike.

(b) We control for special education status using two proxies from state testing data. The first indicates whether a student has previously taken a state test that is intended to be taken by students with disabilities and the second indicates whether a student has previously taken an alternative state test that is intended to be taken by students with an individualized education program. Neither measure exhibits stark racial disparities, though students of color are more likely than white students to have an individualized education program in grades 9–12.

Table 2
Characteristics of students in multi-student fights.

	Grades PK-5			Grades 6-8			Grades 9-12		
	SoC	White	Difference	SoC	White	Difference	SoC	White	Difference
Student attributes									
Always eligible for free/reduced-price lunch?	0.603	0.432	0.172***	0.504	0.314	0.190***	0.460	0.270	0.190***
Sometimes eligible for free/reduced-price lunch?	0.308	0.313	−0.005	0.410	0.421	−0.011	0.464	0.478	−0.014
Never eligible for free/reduced-price lunch?	0.088	0.255	−0.166***	0.086	0.265	−0.179***	0.077	0.252	−0.176***
Disability?	0.412	0.396	0.015	0.007	0.007	−0.000	0.240	0.224	0.016
Individualized education program?	0.419	0.401	0.017	0.104	0.099	0.005	0.367	0.319	0.048**
Level 1 ELA score "not met?"	0.256	0.159	0.096***	0.416	0.291	0.126***	0.353	0.285	0.068***
Level 2 ELA score "nearly met?"	0.119	0.115	0.004	0.231	0.237	−0.006	0.201	0.228	−0.026*
Level 3 ELA score "met?"	0.064	0.115	−0.051***	0.149	0.224	−0.075***	0.120	0.156	−0.036***
Level 4 ELA score "exceeded?"	0.037	0.052	−0.016**	0.035	0.076	−0.040**	0.040	0.081	−0.041***
Missing ELA score?	0.525	0.559	−0.034	0.169	0.173	−0.004	0.286	0.250	0.036*
Level 1 math score "not met?"	0.223	0.117	0.106***	0.475	0.309	0.166***	0.415	0.334	0.081***
Level 2 math score "nearly met?"	0.142	0.136	0.005	0.206	0.260	−0.054***	0.156	0.189	−0.033**
Level 3 math score "met?"	0.083	0.126	−0.043***	0.107	0.163	−0.056***	0.079	0.118	−0.039***
Level 4 math score "exceeded?"	0.031	0.062	−0.031***	0.049	0.099	−0.050***	0.029	0.066	−0.037***
Missing math score?	0.521	0.559	−0.038	0.162	0.169	−0.007	0.321	0.293	0.029
Infraction history									
Fighting infractions last year	0.328	0.262	0.066*	0.301	0.237	0.064***	0.205	0.170	0.035
Other infractions last year	1.121	0.961	0.160	1.431	1.277	0.154	1.512	1.526	−0.015
Fight order (1 = first this year)	1.423	1.339	0.084	1.216	1.188	0.028*	1.091	1.070	0.021*
Any same-race fights last year?	0.061	0.027	0.034***	0.081	0.051	0.030***	0.064	0.049	0.015
Any same-race fights so far this year?	0.126	0.061	0.065***	0.079	0.063	0.017*	0.039	0.023	0.017**
Any multi-race fights last year?	0.029	0.034	−0.005	0.044	0.039	0.006	0.026	0.033	−0.007
Any multi-race fights so far this year?	0.039	0.070	−0.032***	0.040	0.046	−0.006	0.017	0.018	−0.001
Any unknown-race fights last year?	0.121	0.110	0.011	0.114	0.102	0.012	0.085	0.070	0.015
Any unknown-race fights so far this year?	0.120	0.116	0.004	0.068	0.053	0.015**	0.031	0.027	0.004
Observations	2986	1570	4556	4366	3215	7581	2154	1404	3558

Notes: Means of control variables used in Sections 3 and 4 for students of color (SoC) and white students. Exclusionary discipline consists of either a suspension or an expulsion. The sample consists of boys' infractions for "fighting without major injury" (excluding those involving weapons) from multi-student fights that implicate at least two boys for fighting, but do not include girls or implicate other students for non-fighting behaviors. Tests of mean differences are robust to clustering at the school level.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

(c) We control for observed English Language Arts (ELA) and math achievement levels using data from the most recently tested grade. As a general rule, our objective in modeling punishment outcomes is not to control for ability, but rather to control for what an administrator observes (and may consider) when adjudicating misconduct. For this reason, we retain in the sample any students with test scores that are unobservable (e.g., elementary students who have not yet been tested, as tests are not available until the third grade) as they are presumably unobservable to both the econometrician and school administrators. While there are significant racial disparities in achievement within each grade level, the plurality of infractions in our sample are from low-achieving students—this is true for white students and for students of color.

When we control for student attributes in the analyses that follow, we absorb significant differences between students of color and their white peers. Yet, point estimates of the main coefficients of interest are remarkably stable with the addition of controls for student attributes, which suggests that differences in student attributes fail to explain much, if any, of the racial disparities in punishment we observe across the racial composition of fights.

Using up to an additional year of infraction data, we also control for behavior-related attributes that characterize each student's infraction history.

(a) We control separately for the number of fighting infractions and the number of other (non-fighting) infractions from the previous school year to absorb differences in past behavior. Students who select into fights often have an infraction from the previous school year, and students of color tend to have more fighting infractions than their white peers, particularly in grades PK–5 and 6–8.

(b) We control for fight-order fixed effects to absorb differences between first-time and repeat offenses. Most of the infractions in the sample are for a student's first fight of the school year, though there is some evidence that students of color are more likely to be implicated in subsequent fights.

(c) We control for past participation in same-race, multi-race, and unknown-race fights during the previous school year and the current school year to absorb differences in inclinations to participate in multi-race fights. In all three grade spans, students of color are more likely than their white peers to have previously participated in same-race fights. While students of color are less likely to have previously participated in multi-race fights during the current school year in grades PK–5, previous participation in multi-race fights does not exhibit statistically significant racial differences in the other grade spans. Previous participation in fights where the race of other participants is unknown (i.e., fighting infractions without an incident identifier that matches that of another student in the infraction data) does not exhibit significant racial disparities in grades PK–5 and 9–12, though students of color are more likely than white students to have previously participated in an unknown-race fight during the current school year in grades 6–8.

As with student attributes, we absorb significant differences between students of color and their white peers when we control for "infraction history". Under progressive discipline, school administrators assign harsher punishments to students with more extensive disciplinary records, which could, in principle, explain racial disparities in adjudication outcomes. However, we observe little change in the point estimates of the main coefficients of interest when we include infraction history controls, which suggests that differences in past behavior fail to explain the disparities we document across the racial composition of fights.

3. Punishment disparities in school fights

Given the potential for differential selection into incidents (by students, for example) and the potential for differential adjudication of incidents (by vice principals), the difference in the average punishment received by white students and by students of color is not likely capturing the causal relationship of interest—the change in punishment induced by an all-else-equal change in the perception of student race by school officials. For example, if baseline differences in misconduct or punishment vary across schools and there are more students of color in schools with higher baseline levels of misconduct or higher average punishments, then it may well look like students of color are treated more harshly without there ever being any individual actor (e.g., a vice principal) treating students of color differently. Such differences in outcomes are important, but the policy implications can be quite different if no individual school officials are implicated as part of the mechanism that produces differential outcomes.

Below, we consider three increasingly severe margins of punishment and provide estimates of the gap in outcomes for students of color across several specifications. In the end, we will approach a within-incident comparison where we are more inclined to interpret estimates as causal. We will then re-direct our efforts toward identifying other sources of disparate treatment that can explain the advent of race-based differentials in punishment.

3.1. Any exclusionary discipline

In Fig. 1 we begin by reporting unconditioned differences in the adjudication of student misconduct before progressively restricting the variation that identifies racial disparities in punishments. The leftmost estimate for each grade span in Panel A represents the unconditioned racial difference in the probability of receiving any exclusionary discipline. Among the fighting infractions of elementary school students, the probability of receiving any suspension or expulsion is 5.59 percentage points, or 15.2 percent, higher ($p = 0.002$) for students of color than for white students. Relative to the sample standard deviation (σ) of receiving any suspension or expulsion in the estimation sample, this difference corresponds to an effect size on the order of 0.11σ. A significant estimated race differential also exists among the fighting infractions of middle school students (3.96 percentage points, 4.9%, 0.11σ, $p < 0.001$). At 1.52 percentage points (1.7%), the estimated race differential among the infractions of high school students is marginally statistically significant ($p = 0.066$) and relatively small in magnitude (0.05σ).

In Column (2) of each grade span, we control for student attributes (i.e., eligibility for free or reduced-price school meals, past achievement, and proxies for the receipt of special education services), infraction history (i.e., counts of fighting infractions and other infractions from the previous school year, fight-order fixed effects, and indicators for past participation in same-race, multi-race, or unknown-race fights), and school-grade-year fixed effects, absorbing any variation in punishment across schools into the error term for the sample of all fighting infractions. Similar to Kinsler (2011), where a similar specification is estimated using data from North Carolina, this decreases the variation in exclusionary discipline that is attributable to race. However, on other margins of punishment, considered further below, the within-school variation will be suggestive of significant gaps in the adjudication of infractions for students of color compared to white students.

In Column (3) of each grade span, we consider the unconditioned race differential for fighting infractions from schools that report fighting infractions with matching identifiers—it is within these schools that we will have the ability to leverage within-incident variation. As in the full sample, the unconditioned gap in these schools implies that students of color are more likely than white students to receive exclusionary discipline for fighting, though some precision is lost in the elementary and high school samples. The addition of controls and school-grade-year fixed effects in Column (4) also reduces the estimated magnitude of any race differentials.

In columns (5) through (7) we restrict the sample to fighting incidents that explicitly implicate more than one student. For completeness, we again produce estimates of the unconditioned differences and thereafter collapse toward our preferred specification. In columns (6) and (7), for example, we control first for student attributes and then also for infraction history.

In Column (8) of Fig. 1 we control for school heterogeneity with the inclusion of school-grade-year fixed effects, and in columns (11) through (13) we also absorb any unobserved heterogeneity that is specific to incidents. This is where the literature has expressed the most confidence in having retrieved estimates that warrant a causal interpretation. Specifically, we perform these within-incident comparisons by estimating models of the form

$$\mathbb{1}(\text{Punishment} > 0)_{iksgy} = \beta\ \text{SoC}_i + X_i'\Theta + \lambda_{sgy} + \lambda_k + \nu_{iksgy}\ , \tag{1}$$

where $\mathbb{1}(\text{Punishment} > 0)_{iksgy}$ indicates whether a student-infraction i resulted in exclusionary discipline for the student's involvement in incident k.[12] The subscripts s, g, and y index the school, grade, and school year. Incident fixed effects (λ_k) capture unobserved heterogeneity across incidents. Student controls (X_i') adjust for level differences that arise from within-incident variation in student attributes (i.e., eligibility for free or reduced-price school meals, math and reading achievement levels from the previous school year, and proxies for the receipt of special education services) and infraction history (i.e., counts of fighting infractions and other infractions from the previous school year, fight-order fixed effects, and indicators for past participation in same-race, multi-race, or unknown-race fights). In the presence of incident fixed effects, school-grade-year fixed effects (λ_{sgy}) flexibly adjust for level differences that arise from within-incident variation in grade levels, allowing the treatment of students in different grades to vary across schools (e.g., due to differences in grade configuration) and school years (e.g., due to changes in age-specific discipline policies). Our parameter of interest (β) absorbs the average difference in the probability of exclusionary discipline for students of color ($\text{SoC}_i = 1$) relative to white students ($\text{SoC}_i = 0$). The error term (ν_{iksgy}) captures any remaining variation. Throughout the analysis we allow for clustering at the school level.

If students of color are systematically more culpable (e.g., more contributory, or associated systematically with actions that are deemed more severe, or more worthy of punishment), then it would not be surprising to observe punishment differentials that disfavor students of color. This constitutes the assumption that implies a causal interpretation of $\hat{\beta}$—conditional on the full set of controls, school-grade-year fixed effects, and incident fixed effects, we assume that students of color are not differentially culpable on average. If selection into misconduct has school officials being less lenient toward students of color, then estimates of racial gaps in punishment could understate the extent of differential adjudication. That said, the relative severity of fights limits the discretion of teachers in deciding whether to refer students to the principal's office for discipline. We are therefore less concerned that differential selection into incidents explains observed differences, or the lack thereof, in outcomes across race.

We find no statistically significant racial disparity in the probability of receiving any exclusionary discipline within incidents. In the preferred specification, the probability of a suspension or expulsion is 0.99 percentage points higher (2.3%, 0.04σ), on average, for students of color in the elementary grades and 0.84 percentage points higher (1%, 0.04σ), on average, for students of color in grades 6–8, but these estimated differences are indistinguishable from zero at conventional significance levels ($p = 0.254$ in grades PK–5 and $p = 0.235$ in grades 6–8). The estimated within-incident race differential is also

[12] No student has multiple infractions within the same incident.

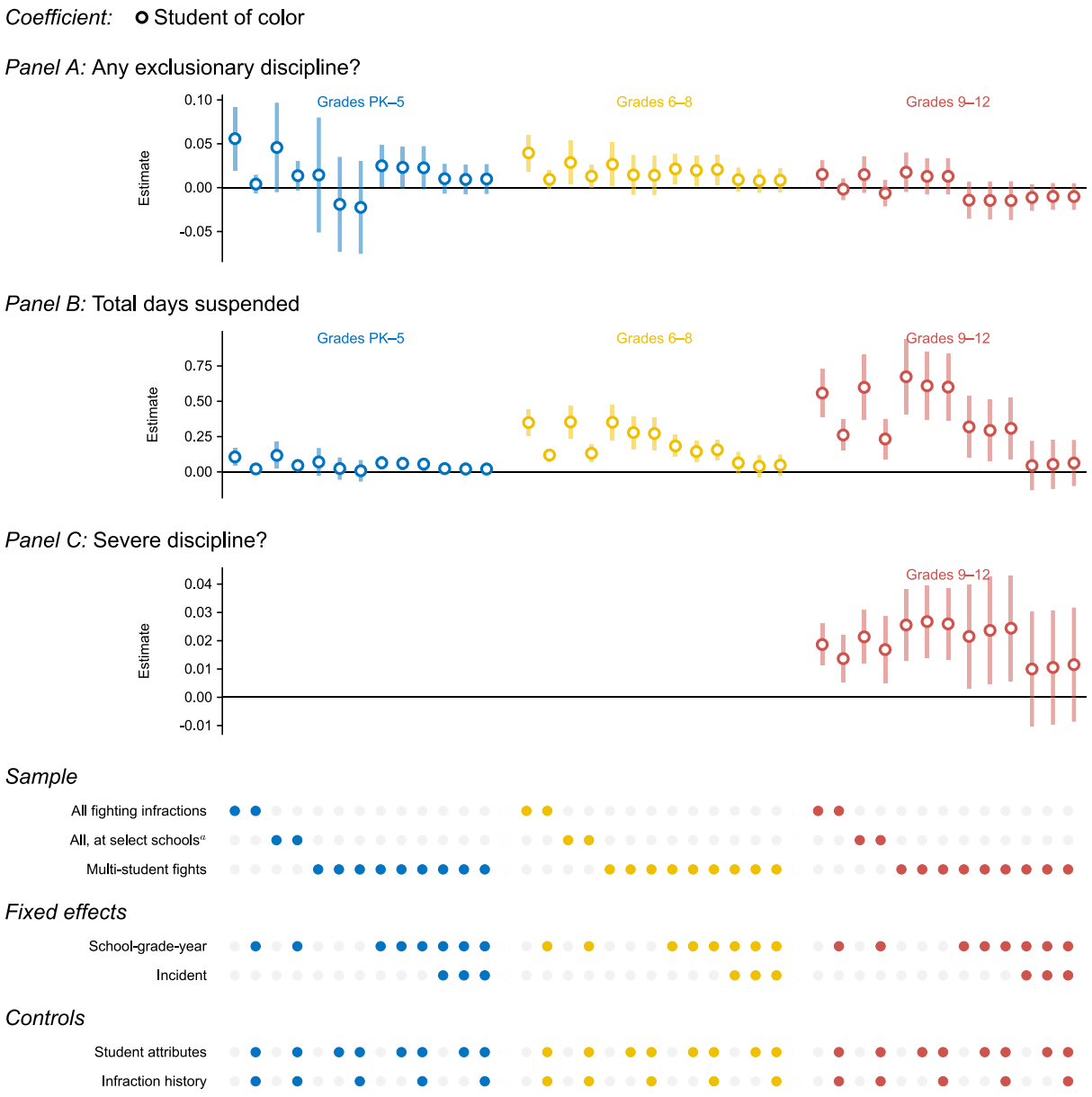

Fig. 1. While students of color receive harsher punishments than white students, on average, differences collapse with the inclusion of incident fixed effects. Notes: Open circles show OLS estimates of racial punishment gaps. Each estimate is from a different regression. The leftmost estimate in each grade span describes a raw punishment gap, and the rightmost estimate describes a within-incident punishment gap from the fully specified model (e.g., see Eq. (1)). The unit of observation is an infraction for "fighting without major injury". The reference category consists of white students' infractions. Solid circles below each set of estimates describe the attributes of each regression: an opaque circle indicates the presence of an attribute and a translucent circle indicates the absence of an attribute. Vertical lines outline 95% confidence intervals adjusted for clustering at the school level. [a]All fighting infractions from schools that report at least one multi-student fight.

indistinguishable from zero in grades 9–12 ($p = 0.191$), though the point estimate is opposite-signed (−1.01 percentage points, −1.1%, -0.07σ). After accounting for unobserved heterogeneity across incidents, race does not appear to influence on whether a student receives exclusionary discipline for fighting.

3.2. Total days suspended

In Panel B of Fig. 1 we perform a similar exercise for the total number of school days a student is suspended for a fighting infraction. Unconditioned, suspensions vary systematically with race. Among the fighting infractions of elementary school students, students of color are suspended for 0.11 days longer, on average, than white students (20.8%, 0.1σ, $p < 0.001$). Large unconditioned gaps are also evident among the fighting infractions of middle school students (0.35 days, 20.7%, 0.2σ, $p < 0.001$) and high school students (0.56 days, 19.2%, 0.23σ, $p < 0.001$).

In contrast to Panel A, the addition of controls and school-by-year fixed effects fails to explain away statistically significant racial disparities in suspension lengths suspended in any of the grade spans or samples considered.[13] For example, estimates in Column (8) imply that high school students of color are, on average, suspended for 0.32 days longer (10.9%, 0.21σ, $p = 0.005$) than white high school students implicated for fighting in the same school, grade, and school year. Similar disparities exist in multi-student fights among middle school students (0.18 days, 10.7%, 0.15σ, $p < 0.001$) and elementary school

[13] The estimate for grades PK–5 in Column (2) is marginally statistically significant ($p = 0.063$).

students (0.06 days, 11.9%, 0.14σ, $p = 0.005$), and estimates are robust to the inclusion of controls for student attributes and infraction history in columns (9) and (10).

However, when we absorb incident-specific heterogeneity in columns (11) through (13), there is less evidence of racial disparities. Conditional on incident fixed effects, school-by-year fixed effects, and the full set of controls, the estimated racial disparities in suspension lengths are small in magnitude and statistically indistinguishable from zero among the fighting infractions of elementary school students (0.02 days, 3.8%, 0.05σ, $p = 0.377$), middle school students (0.05 days, 2.8%, 0.04σ, $p = 0.213$), and high school students (0.06 days, 2.2%, 0.04σ, $p = 0.448$). Within-incident variation does not support the claim that there are significant racial disparities in the amount of time students are suspended for fighting.

3.3. Severe discipline

In Panel C of Fig. 1 we examine racial disparities in the probability of receiving severe discipline, which we define as suspensions longer than 10 school days (roughly two weeks of instruction) or expulsions. As in Table 1, severe discipline is rare, and small cell sizes prevent us from reporting the results of within-incident comparisons for elementary and middle school students. We therefore restrict our attention to the fighting infractions of high school students, in which severe outcomes are rare but not rare enough to impede the analysis.

As with the other margins of punishment, we estimate a significant unconditioned difference in the probability of severe discipline: students of color are 1.87 percentage points, or 100 percent, more likely than white students to be suspended for more than 10 school days or expelled as a consequence of a fighting infraction (0.11σ, $p < 0.001$). Moreover, the addition of school-grade-year fixed effects and controls for student attributes and infraction history fails to reduce the magnitude of the estimated race differential. For example, the estimate in Column (10) suggests that conditional on the full set of controls and school-grade-year fixed effects, students of color are 2.44 percentage points (263.7%) more likely than white students to receive severe discipline (0.21σ, p = 0.01).

Consistent with the results in panels A and B, the addition of incident fixed effects reduces the estimated influence of race on the probability of receiving severe discipline. In the preferred specification, the probability of severe discipline is 1.15 percentage points (124.6%, 0.1σ) higher for students of color, on average, but the estimated difference is indistinguishable from zero at conventional significance levels ($p = 0.26$). Though the difference in probability is statistically insignificant, we cannot rule out meaningful effect sizes at the upper bound of the 95-percent confidence interval.

Still, after accounting for unobserved incident-specific heterogeneity, no margin of punishment supports that there are statistically significant differences in the disciplinary actions imposed on students of color—this is true for elementary, middle, and high school students. Although some estimates have relatively wide confidence intervals—namely those of suspension length and severe discipline gaps in high school—others are precise zeros, giving us an additional degree of confidence that the adjudication of the infractions of students of color is not systematically different from that of white students implicated in the same fight. If white students and students of color are equally culpable, on average, for their involvement in a fight, then differential treatment within incidents explains very little of the aggregate racial disparities.

4. Looking across incidents to identify where gaps arise

We have demonstrated that the difference in the aggregate experience of students of color can be made indistinguishable from zero upon the inclusion of observable attributes.[14] However, the identifying variation also changes across columns of Fig. 1, and we need not be identifying the same parameter at all—we potentially learn about a different parameter that informs the research question differently. For example, the inclusion of school fixed effects is typically justified with an appeal to absorbing things about schools that lead to punishments being different—maybe there are more serious fights in some schools than in others, or teachers with different tolerances in some schools than in others. So, with the inclusion of school (or school-grade-year) fixed effects, we are able to abstract away from such differences and identify whether there are remaining differences in the treatment of students of color by the decision makers within those school-grade-years—and there are. In arriving at that parameter, then, all of the punishments allocated by administrators at a school to students in a grade are pooled together to identify how students of color are treated relative to white students at the same school and grade level during the same school year. (It is the average difference within school-grade-years that is then reported, of course.)

However, the inclusion of *incident* fixed effects operates in a fundamentally different way, and changes what is being identified in ways that are important to consider—it is not merely a further step toward better identifying the same parameter. Unlike the inclusion of school-grade-year fixed effects, the inclusion of incident fixed effects separates the decisions of individual administrators into two different categories, with only one of them used to identify the differential. Specifically, with the inclusion of incident fixed effects, same-race fights contribute nothing to identifying the punishment differentials experienced by students of color. Further, as it is there where we see punishment gaps collapse, this suggests that there is possibly an important contextual change in moving from school-grade-year fixed effects models in columns (8)–(10) of Fig. 1 to incident fixed effects models in columns (11)–(13). In short, the parameter identified in the existing literature does not fully characterize the relevant experiences of students of color, insofar as the within-incident differentials capture only the actions taken when students of color are implicated with white students. And there, we will beg to differ with what parameter is being identified.

The collapse of punishment disparities within incidents does not necessarily imply that students of color are being treated equally. For example, one potential explanation for the absence of significant gaps in punishments across students in multi-race fights is that within-incident variation in race offers a degree of salience that induces more equal treatment of students of color. That is, while adjudicating an incident, that one student is white and the other is a student of color may bring a beneficial awareness to the need to guard against implicit biases. Alternatively, it could be that explicit biases are more costly to act on within incidents. For example, while harsher punishments across incidents could be justified by an appeal to some *fights* being worse than others, similar claims may be unavailable when justifying systematically harsher punishments for some *races* over others. If punishment gaps were robust to the inclusion of incident fixed effects, it would be consistent with administrators believing that students of particular races are worse on average or otherwise deserving of harsher punishment. Aversion to publicly displaying such a belief may further encourage more equal treatment within incidents.

That punishment gaps attenuate when we identify off of within-incident variation is also consistent with gaps initially having been driven by race-based differences in the inclination of parents to advocate for their children, or for their advocacy to exert varying degrees of influence on punishments. For a differential-advocacy story to explain that punishments are harsher for students of color, generally, but equal

[14] This is true of the literature more generally, insofar as the inclusion of observable attributes attenuates estimated racial disparities in some settings (e.g., Barrett et al., 2021; Shi & Zhu, 2022) and eliminates them in others (e.g., Kinsler, 2011).

within incidents, it would need to be the case that administrators extend the benefits of racially disparate advocacy to others involved in the same fight, regardless of their race. Whatever the specific mechanism, our analysis suggests that administrators are better able to maintain equality norms within fights than they are across fights.

4.1. Specification

By absorbing the unobserved heterogeneity associated with specific incidents into the error term, the analysis in Fig. 1 identifies only those factors that vary within incidents—we necessarily lose the context that would come from the comparison of multi-race fights alongside same-race fights, where some of the mechanisms that induce equal treatment are absent. In Fig. 2 we therefore consider the punishments of students of color across multi-student fights—dropping the incident fixed effects from the earlier analysis allows for the comparison of multi-race and same-race fights.[15] Specifically, we estimate models of the form

$$\text{Punishment}_{iksgy} = \beta\ \text{SoC}_i + \tau\ \text{Multiracial}_k + \phi\ \text{SoC}_i \times \text{Multiracial}_k + X_i'\Theta + \lambda_{sgy} + \nu_{iksgy}\ , \quad (2)$$

where $\text{Punishment}_{iksgy}$ is the disciplinary intervention assigned to student-infraction i for the student's involvement in incident k while they were enrolled in school s for grade g during school year y. As before, we control for student attributes and infraction history and identify racial gaps (β) using within-school-grade-year variation, though now β characterizes gaps across same-race fights (as opposed to gaps within multi-race fights). As selection into multi-race fights may differ, we absorb any level effect associated with multi-race fights in τ. However, our interest is in how that treatment changes for students of color across the changing racial composition of fights (ϕ). In a way, we are asking whether being implicated with a white student induces changes in the punishments assigned to students of color.[16]

For $\hat{\phi}$ to identify how school administrators treat students of color differently in the presence of a white student, we must assume that selection into multi-race fights does not depend on unobserved race-specific differences in behavior. By estimating level differences in the adjudication of multi-race versus same-race fights (i.e., $\hat{\tau}$), we absorb differences in the severity of multi-race fights that are common to students of color and white students.[17] With this in mind, our assumption about selection into multi-race fights requires that if there is a change in aggression displayed by students of color when fighting white students it must be similar to the change in aggression displayed by white students when fighting students of color.

For $\hat{\beta}$ to identify racial bias in the adjudication of same-race fights, we must also assume that students of color and white students are equally culpable for their behavior in same-race fights after conditioning on observable student attributes, infraction histories, and school-grade-year fixed effects. If the equal culpability assumption holds, then $\hat{\phi}$ identifies the effect of variation in race within an incident on racial *bias* in adjudication outcomes. Without the equal culpability assumption, $\hat{\phi}$ identifies the effect of variation in race on *disparities* in adjudication outcomes—a weaker, but still policy-relevant interpretation.

4.2. Results

Having estimated several models based on Eq. (2), we plot two coefficient estimates from each model in Fig. 2. The first is the estimated difference in outcomes for students of color. This difference is identified off of same-race fights, so it reflects the average difference in outcomes across fights that implicated only students of color and fights that implicated only white students. The second is the estimated difference-in-differences for students of color in fights that also implicated a white student.

In Panel A of Fig. 2 we consider racial disparities in the probability of receiving any exclusionary discipline as a consequence of a fighting infractions for students in each grade span. Overall, we find that students of color are only more likely to experience exclusionary discipline when they are implicated with only other students of color. This is most evident among the fights of middle school students, where our preferred specifications suggest that (i) students of color *in fights that only implicate students of color* experience significantly higher rates of exclusionary discipline (4.83 percentage points, 0.23σ, $p = 0.007$) and (ii) the increase in the probability of exclusionary discipline for students of color is offset when there is a white student implicated in the same fight (−4.7 percentage points, -0.22σ, $p = 0.011$). The sum of those coefficients—that is, the marginal effect of being a student of color in a multi-race fight—is indistinguishable from zero (0.13 percentage points, 0.01σ, $p = 0.86$), which is consistent with the presence of a white student *fully* offsetting the average difference in exclusionary discipline. The same pattern is also evident among fights in grades PK–5, where students of color experience a higher probability of exclusionary discipline when they are implicated with only students of color (5.31 percentage points, 11.7%, 0.23σ, $p = 0.18$) than when they are implicated with white students (1.21 percentage points, 2.6%, 0.05σ, $p = 0.214$), though precision falls short of conventional significance levels. In grades 9–12, however, we do not observe significant differences for students of color when they are implicated with only students of color (−2.35 percentage points, −2.6%, -0.15σ, $p = 0.247$) or when they are implicated with white students (−0.94 percentage points, −1.1%, -0.06σ, $p = 0.32$). Across all three grade spans, inferences are robust to restricting the sample to fights that involve exactly two students.[18]

In Panel B we consider differences in suspension length, where we find similar patterns, and with enough precision to suggest that differences in the treatment of students across the racial composition of fights contribute to racial disparities within schools. Relative to white students in same-race fights, students of color in fights that implicate only other students of color are suspended for 0.16 days longer, on average, in elementary school (30.2%, 0.37σ, $p = 0.019$), 0.32 days longer in middle school (19.6%, 0.27σ, $p < 0.001$), and 0.69 days longer in high school (25.1%, 0.45σ, $p < 0.001$). However, when implicated with white students, students of color are suspended for no longer, on average, than are white students in all-white fights—this is true among elementary school students (0.02 days, 4.5%, 0.06σ, $p = 0.268$), middle school students (0.03 days, 2.3%, 0.03σ, $p = 0.345$), and high school students (0.04 days, 1.3%, 0.02σ, $p = 0.735$). As in Panel A, this pattern is robust to restricting the sample to fights that involve exactly two students.

[15] Multi-race fights make up 34.7 percent of multi-student fights while 42.5 percent implicate only students of color and the remaining 22.7 percent implicate only white students.

[16] Here, ϕ captures a different relationship than what would be identified from a model with student fixed effects, as conditioning on student fixed effects would restrict the identifying variation to students with repeat infractions that vary in their racial composition. As with the inclusion of incident fixed effects, the disciplinary experiences of those who contribute identifying variation in the presence of student fixed effects may not necessarily reflect the experiences of the broader set of students who receive infractions. Indeed, in our data, only 3.08 percent of student-grade observations include both same-race and multi-race fights.

[17] Point estimates of τ are generally positive, capturing that multi-race fights tend to be punished more heavily than same-race fights.

[18] Given that most of the fights in our data are small, it is not surprising that the results are robust to restricting the sample to two-student fights. Pooling across grade spans, 92.2 percent of multi-student fights involve only two students, 5.3 percent involve three students, 1.6 percent involve four students, 0.5 percent involve five students, 0.3 percent involve six-to-eight students, and precisely 0 percent involve nine or more students. While results are insensitive to the inclusion or exclusion of larger fights, we believe that fights involving exactly two students provide the cleanest environment for identifying racial disparities across the racial composition of fights, so we upweight them in our discussion.

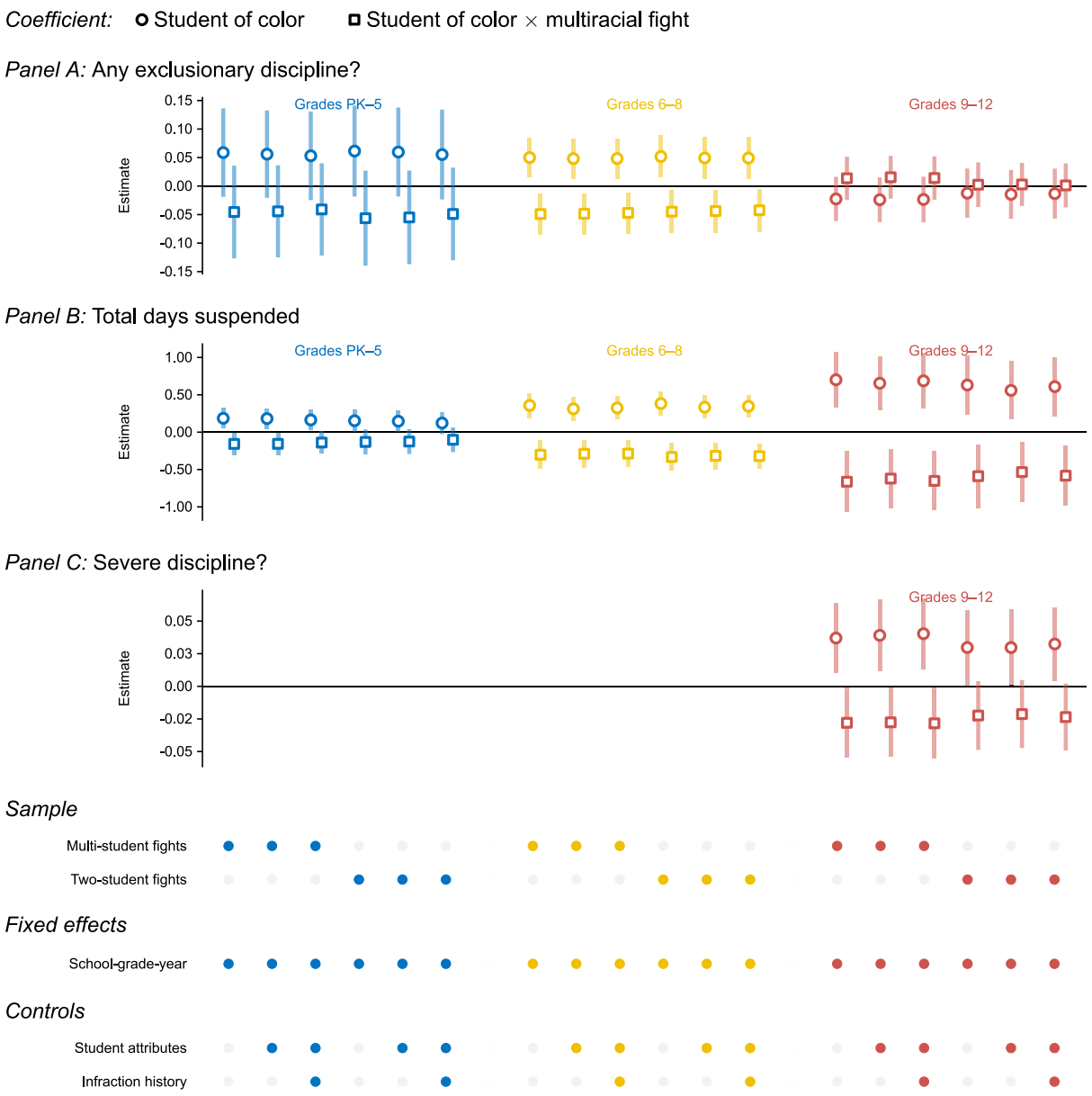

Fig. 2. Do incident fixed effects "explain away" important sources of race-based differences in punishment? Across-incident variation suggests that the punishment of students of color depends on the presence of a white student. Notes: Open circles and squares show OLS estimates of coefficients from Eq. (2). Each set of two estimates is from a different regression. The unit of observation is an infraction, and the sample consists of infractions from multi-student fights in which all students receive infractions for "fighting without major injury." The reference category consists of white students' infractions from all-white fights. Solid circles below each set of estimates describe the attributes of each regression: an opaque circle indicates the presence of an attribute and a translucent circle indicates the absence of an attribute. Vertical lines outline 95% confidence intervals adjusted for potential clustering at the school level.

In Panel C we consider differences in the probability of severe discipline among high school students. Relative to white students in same-race fights, students of color in "same-race" fights are more likely to experience severe discipline (4.03 percentage points, 0.34σ, $p = 0.004$) whereas students of color in multi-race fights are not (1.19 percentage points, 0.1σ, $p = 0.218$). The gap in severe discipline across same-race fights persists when we restrict the sample to fights that involve exactly two students (3.24 percentage points, 0.32σ, $p = 0.024$), as does the offsetting difference for being implicated with a white student (−2.36 percentage points, -0.23σ, $p = 0.071$). As in the broader sample of multi-student fights, students of color are no more likely to be expelled when they are implicated with white students (0.88 percentage points, 0.09σ, $p = 0.389$).[19]

The offsetting differences in Fig. 2 suggest that the within-school disparities in Fig. 1 are driven by differences in punishment *across* same-race fights. When implicated in fights with at least one white student, students of color are punished no differently, on average,

[19] We suppress "impact" estimates (percentage changes over the mean of the reference group) because severe discipline is an exceedingly rare outcome for white students in same-race fights.

than white students implicated for fighting in the same school, grade, and school year. When implicated in fights without white students, however, students of color receive systematically harsher punishments than those imposed on white students.

Depending on the grade span and margin of punishment, the estimated disparity in punishment between students of color and white students in same-race fights is often large—in several cases the point estimate is nearly identical to the unconditioned race gap. Indeed, some of the race differentials that we estimate across same-race fights are larger in magnitude than within-school gaps documented elsewhere in the literature (e.g., Anderson & Ritter, 2020; Barrett et al., 2021; Kinsler, 2011; Shi & Zhu, 2022).

To explain away the variation we observe in the data (conditioning on school-grade-year fixed effects, student characteristics, infraction history, and the racial composition of fights) one would have to assume that (i) students of color who select into fights with only other students of color are somehow more deserving of punishment than white students who select into fights with only other white students, and (ii) students of color who select into fights with white students are somehow less deserving of punishment than students of color who select into fights with only other students of color. Having conditioned the sample on a narrow band of behavior, we doubt that differential selection among students of color into fights with white students explains the collapse of punishment disparities within fights.

If students of color and white students in same-race fights are equally culpable (after conditioning on school-grade-year fixed effects and the full set of controls), then the patterns we document are consistent with disparate treatment of all-white fights and fights involving only students of color. The full characterization of the data-generating process—with within-incident variation coming from multi-race fights—strongly suggests that the presence of a white student moves administrators toward equal treatment, consistent with administrators correcting biases when racial differences are more salient.

5. Disparities for black and hispanic students

In this section we estimate punishment disparities separately for Black and for Hispanic students, the two largest groups of students of color represented in our data. To do so, however, we face a tradeoff—characterizing disparities for specific racial groups will necessitate that we no longer estimate disparities separately by grade span.

With that restriction in mind, in Fig. 3 we report estimates of Black-white and Hispanic-white punishment disparities in school fights. To identify Black-white gaps we restrict the sample to fights that involve only Black or white students, and to identify Hispanic-white gaps we restrict the sample to fights that involve only Hispanic or white students. As such, the infractions of white students in all-white fights are included in both analyses (3440 student-infractions), and the infractions of Black (Hispanic) students who were implicated with other non-Black (non-Hispanic) students of color are excluded (1140 Black and 1112 Hispanic student-infractions). The Black-white analysis includes a total of 5659 student-infractions (= 1672 Black + 3987 white) and the Hispanic-white analysis includes a total of 7633 student-infractions (= 2995 Hispanic + 4638 white).[20] To bridge the analysis of separate disparities with the analysis presented in Fig. 1, we also report pooled estimates of disparities for students of color.

Overall, confidence intervals tend to be wider for estimates of Black-white and Hispanic-white gaps than for estimates of disparities for students of color more broadly, which is consistent with our *ex ante* concerns about power. However, the story in Fig. 3 is largely the same as the analysis that pools races (i.e., Fig. 1). The main exception is that disaggregating race reveals a small, marginally significant within-incident gap in the probability of any exclusionary discipline for Hispanic students (1.2 percentage points, 1.5%, 0.07σ, $p = 0.052$).

Where we can make comparisons to the literature we note that our point estimates of Black-white and Hispanic-white disparities within fights tend to be smaller. For example, Liu et al. (2024) finds that Black students in a large California school district are 3.4 percentage points more likely to receive exclusionary discipline than white students implicated in the same violent incident, whereas we find that Black students are no more likely to receive exclusionary discipline (0.38 percentage points, 0.5%, 0.02σ, $p = 0.734$). Liu et al. (2024) also finds that Hispanic students are 3 percentage points more likely to receive exclusionary discipline, which is less than our point estimate of 1.2 percentage points. Among students implicated in the same fight, Barrett et al. (2021) finds that Black students in Louisiana suspended for 0.05 days longer, on average. Our estimate of the Black-white disparity (0.03 days, 1.7%, 0.03σ) is similar in magnitude to that of Barrett et al. (2021), but statistically indistinguishable from zero ($p = 0.588$).

Where we depart most from the literature is in setting the within-fight experiences of students of color within the context of differences in the treatment of students of color *across* the racial composition of fights. Following the analysis in Section 4, in Fig. 4 we illustrate how Black-white and Hispanic-white disparities depend on the presence of a white student in a fight. The estimates in Panel A suggest that Black students in all-Black fights are more likely to receive exclusionary discipline than white students in all-white fights (7.09 percentage points, 9%, 0.4σ, $p = 0.012$), but Black students in fights with white students are not (0.31 percentage points, 0.4%, 0.02σ, $p = 0.783$). Hispanic-white disparities in exclusionary discipline exhibit a similar pattern, though point estimates are smaller in magnitude and statistically indistinguishable from zero. The estimates in Panel B suggest that Black students are suspended for 0.45 days longer (27.5%, 0.49σ, $p = 0.009$), on average, but only in all-Black fights—in fights with white students, the implied disparity is statistically indistinguishable from zero (0.04 days, 2.8%, 0.05σ, $p = 0.423$). Likewise, we find that Hispanic students are suspended for 0.34 days longer (20.6%, 0.34σ, $p = 0.003$), on average, but only in all-Hispanic fights—in fights with white students, the implied disparity also collapses to zero (0.02 days, 1.4%, 0.02σ, $p = 0.623$). The estimates in Panel C exhibit similar patterns, though the implied Hispanic-white disparities are smaller in magnitude and statistically insignificant. As in Fig. 2, variation in race within an incident appears to move administrators toward equal treatment. The results in Fig. 4 also suggest that this movement is especially pronounced for Black students.

6. Conclusion

Racial disparities in the incidence of exclusionary discipline have increased since race-based gaps in suspensions were first documented (Children's Defense Fund, 1975; Losen, Hodson, Keith II, Morrison, & Belway, 2015). In an effort to reduce discipline gaps, policymakers have begun to roll back strict "zero tolerance" discipline policies that have been shown to have a disparate impact on students of color (Curran, 2016). For example, some school districts have implemented policies that mandate the elimination of exclusionary interventions for low-level offenses (Craig & Martin, 2023; Lacoe & Steinberg, 2018; Pope & Zuo, 2023; Steinberg & Lacoe, 2018), while others have experimented with less punitive disciplinary interventions, such as restorative justice (Glenn, Barrett, & Lightfoot, 2021). However, the ultimate success of any disciplinary reform depends, in part, on the ability of school officials to enforce policies impartially. With evidence that educators hold biases that disfavor students of color (Chin, Quinn, Dhaliwal, & Lovison, 2020), the extent to which those biases manifest in disparate treatment can have important implications for the effectiveness of education reforms, including those concerning the use of discipline.

We consider the potential for racial bias in disciplinary outcomes by comparing the punishments of students within and across fights.

[20] A similarly constructed Asian-white analysis would include only 228 infractions by Asian students.

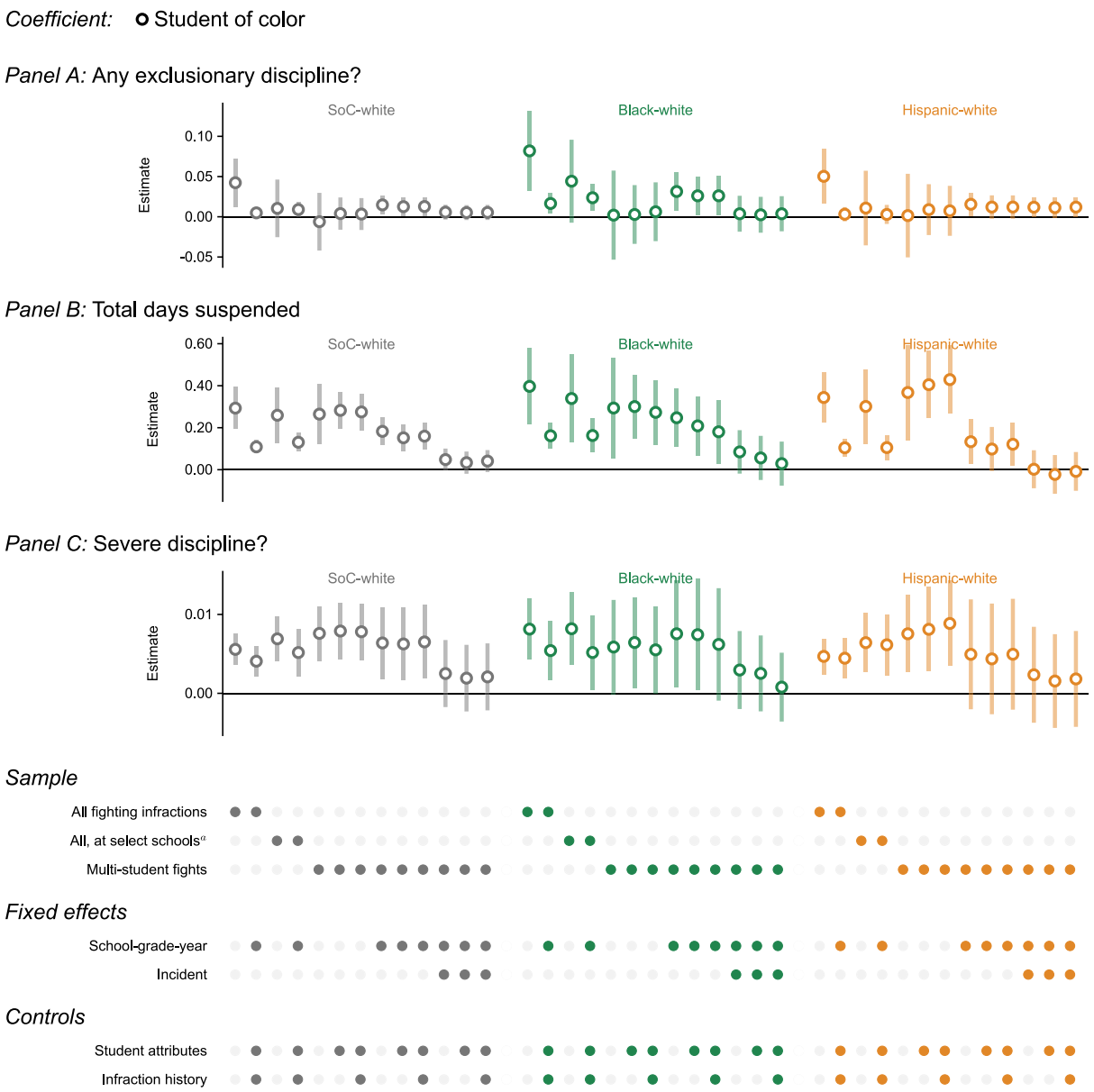

Fig. 3. Identifying within-incident disparities (as in Fig. 1) separately for Black and Hispanic students. Notes: Open circles show OLS estimates of racial punishment gaps. Each estimate is from a different regression. The leftmost estimate in each grade span describes a raw punishment gap, and the rightmost estimate describes a within-incident punishment gap from the fully specified model (e.g., see Eq. (1)). The unit of observation is an infraction for "fighting without major injury". The reference category consists of white students' infractions. The analysis of Black-white punishment disparities (i) restricts the sample of infractions to those of Black and white students and (ii) restricts the sample of multi-student fights to those that include Black students only, white students only, or both Black and white students together without students from other backgrounds. The analysis of Hispanic-white punishment disparities (i) restricts the sample of infractions to those of Hispanic and white students and (ii) restricts the sample of multi-student fights to those that include Hispanic students only, white students only, or both Hispanic and white students together without students from other backgrounds. Solid circles below each set of estimates describe the attributes of each regression: an opaque circle indicates the presence of an attribute and a translucent circle indicates the absence of an attribute. Vertical lines outline 95% confidence intervals adjusted for clustering at the school level. [a]All fighting infractions from schools that report at least one multi-student fight.

Consistent with existing evidence (Anderson & Ritter, 2020; Barrett et al., 2021; Beck & Muschkin, 2012; Gopalan & Nelson, 2019; Liu et al., 2024; Shi & Zhu, 2022), we document significant within-school disparities in adjudication outcomes. However, we find little evidence that students of color receive systematically harsher punishments than white students implicated in the same incident. Point estimates on within-incident disparities are small and statistically indistinguishable from zero. Across-incident estimates suggest that within-school disparities are instead driven by the tendency for fights involving only students of color to elicit significantly harsher punishments than fights involving only white students. Moreover, variation in race within an incident appears to offset within-school punishment differentials for students of color, as students of color in multi-race fights receive punishments that are statistically indistinguishable from those assigned to white students in all-white fights.

We find encouragement insofar as the data-generating process supports that biases are correctable where race and equality norms are more salient. That being said, our results imply that purging all within-incident disparities in punishment would do little to close the gap in disciplinary outcomes between students of color and their white peers. Our results also raise questions about the prospects of school accountability measures that leverage incident identifiers to detect differential treatment—relying on within-incident comparisons to monitor unequal treatment would falsely signal an equality in outcomes, understating the extent of differential treatment.

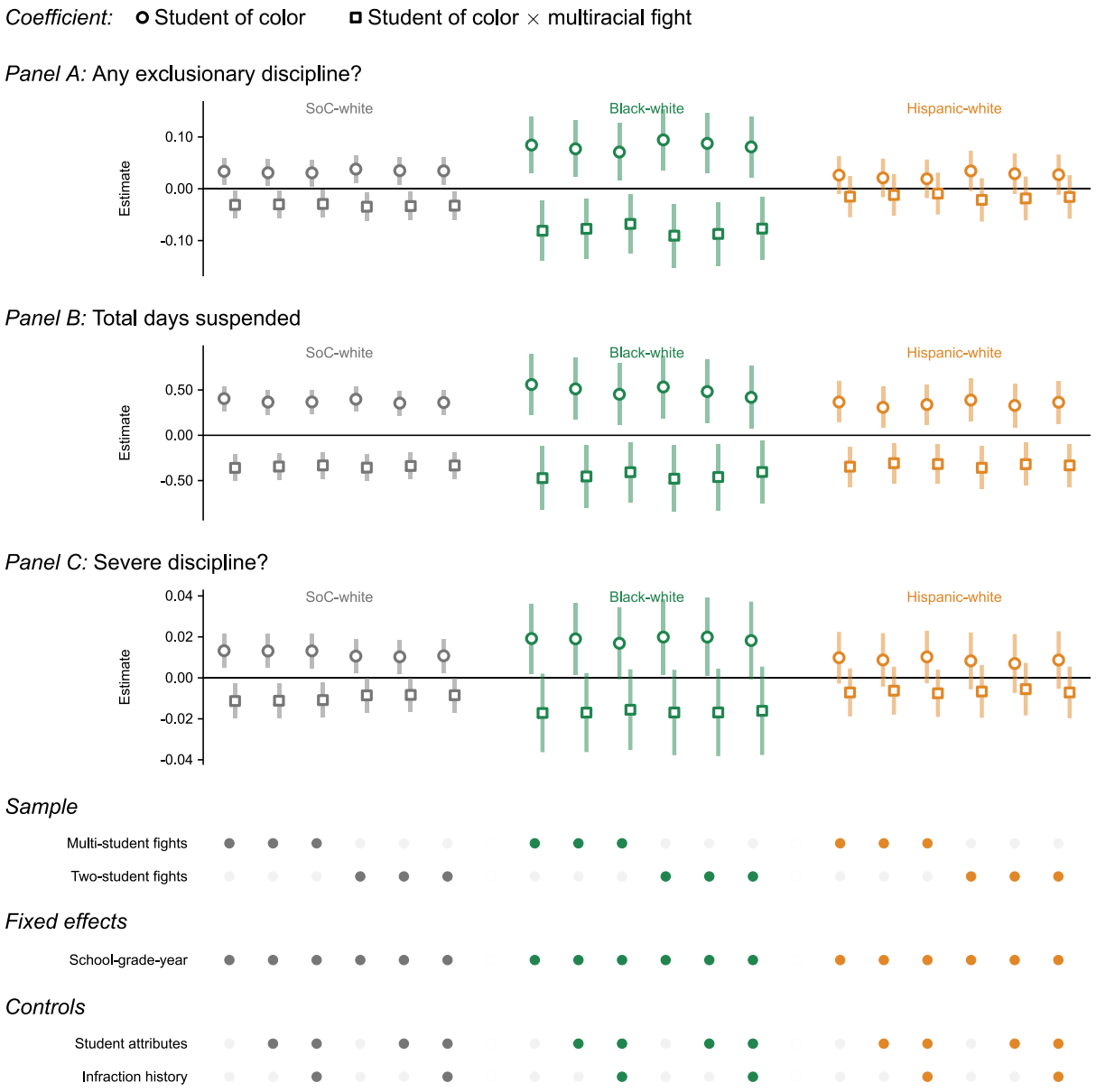

Fig. 4. Identifying across-incident disparities (as in Fig. 2) separately for Black and Hispanic students. Notes: Open circles and squares show OLS estimates of coefficients from Eq. (2). Each set of two estimates is from a different regression. The unit of observation is an infraction, and the sample consists of infractions from multi-student fights in which all students receive infractions for "fighting without major injury". The reference category consists of white students' infractions from all-white fights. The analysis of Black-white punishment disparities restricts the sample to fights that include Black students only, white students only, or both Black and white students together without students from other backgrounds. The analysis of Hispanic-white punishment disparities restricts the sample to fights that include Hispanic students only, white students only, or both Hispanic and white students together without students from other backgrounds. Solid circles below each set of estimates describe the attributes of each regression: an opaque circle indicates the presence of an attribute and a translucent circle indicates the absence of an attribute. Vertical lines outline 95% confidence intervals adjusted for potential clustering at the school level.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Anderson, K. P., & Ritter, G. W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a US state. *Education Policy Analysis Archives*, *25*(9).

Anderson, K. P., & Ritter, G. W. (2020). Do school discipline policies treat students fairly? Evidence from Arkansas. *Educational Policy*, *34*(5), 707–734.

Anderson, K. P., Ritter, G. W., & Zamarro, G. (2019). Understanding a vicious cycle: The relationship between student discipline and student academic outcomes. *Educational Researcher*, *48*(5), 251–262.

Arcidiacono, P., Aucejo, E. M., & Spenner, K. (2012). What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice. *IZA Journal of Labor Economics*, *1*(1).

Bacher-Hicks, A., Billings, S. B., & Deming, D. J. (2019). *The school to prison pipeline: Long-run impacts of school suspensions on adult crime*. (Working Paper No. 26257). National Bureau of Economic Research, https://www.nber.org/papers/w26257.

Barrett, N., McEachin, A., Mills, J. N., & Valant, J. (2021). Disparities and discrimination in student discipline by race and family income. *Journal of Human Resources*, *56*(3), 711–748.

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior and Organization*, *39*(1), 41–58.

Beck, A. N., & Muschkin, C. G. (2012). The enduring impact of race: Understanding disparities in student disciplinary infractions and achievement. *Sociological Perspectives*, *55*(4), 637–662.

Card, D., & Payne, A. A. (2021). High school choices and the gender gap in STEM. *Economic Inquiry*, *59*(1), 9–28.

Carrell, S. E., & Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, *2*(1), 211–228.

Chetty, R., Dobbie, W., Goldman, B., Porter, S. R., & Yang, C. S. (2024). *Changing opportunity: Sociological mechanisms underlying growing class gaps and shrinking race gaps in economic mobility*. (Working Paper No. 32697). National Bureau of Economic Research, https://www.nber.org/papers/w32697.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the United States: An intergenerational perspective. *Quarterly Journal of Economics*, *135*(2), 711–783.

Children's Defense Fund (1975). *School suspensions: Are they helping children?*. Cambridge, MA: Children's Defense Fund.

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educational Researcher*, *49*(8), 566–578.

Craig, A. C., & Martin, D. C. (2023). *Discipline reform, school culture, and student achievement*. (Discussion Paper No. 15906). Institute of Labor Economics (IZA), https://www.iza.org/publications/dp/15906/.

Curran, F. C. (2016). Estimating the effect of state zero tolerance laws on exclusionary discipline, racial discipline gaps, and student behavior. *Educational Evaluation and Policy Analysis*, *38*(4), 647–668.

Fudenberg, D., & Tirole, J. (1986). A "signal-jamming" theory of predation. *Rand Journal of Economics*, *617*(3), 366–376.

Glenn, B., Barrett, N., & Lightfoot, E. (2021). *The effects and implementation of restorative practices for discipline in New Orleans schools*. Technical report. Education Research Alliance for New Orleans, https://educationresearchalliancenola.org/files/publications/Restorative-Approaches-Technical-Report.pdf.

Gopalan, M., & Nelson, A. A. (2019). Understanding the racial discipline gap in schools. *AERA Open*, *5*(2).

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies*, *66*(1), 169–182.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576.

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

Kinsler, J. (2013). School discipline: A source or salve for the racial achievement gap? *International Economic Review*, *54*(1), 355–383.

Lacoe, J., & Steinberg, M. P. (2018). Rolling back zero tolerance: The effect of discipline policy reform on suspension usage and student outcomes. *Peabody Journal of Education*, *93*(2), 207–227.

Liu, J., Hayes, M. S., & Gershenson, S. (2024). From referrals to suspensions: New evidence on racial disparities in exclusionary discipline. *Journal of Urban Economics*, *141*, Article 103453.

Losen, D. J., Hodson, C. L., Keith II, M. A., Morrison, K., & Belway, S. (2015). *Are we closing the school discipline gap?*. Los Angeles: The Center for Civil Rights Remedies, University of California.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open*, *3*(1).

Pope, N. G., & Zuo, G. W. (2023). Suspending suspensions: The education production consequences of school suspension policies. *The Economic Journal*, *133*(653), 2025–2054.

Reykdal, C., Weaver-Randall, K., & Ireland, L. (2018). *Comprehensive education data and research system (CEDARS) appendix manual, version 10.2*. Washington State Office of the Superintendent of Public Instruction.

Ritter, G. W., & Anderson, K. P. (2018). Examining disparities in student discipline: Mapping inequities from infractions to consequences. *Peabody Journal of Education*, *93*(2), 161–173.

Shi, Y., & Zhu, M. (2022). Equal time for equal crime? Racial bias in school discipline. *Economics of Education Review*, *88*, Article 102256.

Smith, C. W., Kreitzer, R. J., & Suo, F. (2020). The dynamics of racial resentment across the 50 US states. *Perspectives on Politics*, *18*(2), 527–538.

Sorensen, L. C., Bushway, S. D., & Gifford, E. J. (2022). Getting tough? The effects of discretionary principal discipline on student outcomes. *Education Finance and Policy*, *17*(2), 255–284.

Steinberg, M. P., & Lacoe, J. (2018). Reforming school discipline: School-level policy implementation and the consequences for suspended students and their peers. *American Journal of Education*, *125*(1), 29–77.

Welsh, R. O., & Little, S. (2018). The school discipline dilemma: A comprehensive review of disparities and alternative approaches. *Review of Educational Research*, *88*(5), 752–794.