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

Kyle Raze † Glen R. Waddell ‡

August 2022

Abstract

Racial gaps in the adjudication of student misconduct are well documented—for similar behaviors, students of color are more likely to be disciplined and discipline tends to be harsher. While students of color do receive harsher punishments, on average, we show that this differential depends on the racial composition of incidents. Consistent with administrators moving toward equal treatment when variation in race is more salient, multi-race incidents evidence no differentials in our data. In fact, when a white student is implicated in the same incident as a student of color, punishments imposed on students of color are indistinguishable from those imposed on white students in all-white incidents.

Keywords: student behavior, school discipline, racial discrimination

JEL Classification: I24, J71

* We thank Jon Davis, Seth Gershenson, Zack Hawley, Mike Kuhn, and David Liebowitz for their insightful comments.

† Doctoral candidate at the University of Oregon, raze@uoregon.edu.

‡ Professor of Economics at the University of Oregon and Research Fellow at IZA Bonn, waddell@uoregon.edu.

1 Introduction

With an extensive literature identifying racial disparities in many outcomes, any degree of hysteresis in the production of racial disparities suggests that there are gains to correcting early differences in the experiences of those of different racial backgrounds. In this way, school environments are an important setting to consider—it is in these formative years that students are making human capital investment decisions and forming expectations of their own comparative advantages and relative strengths.[1] In this paper, we examine patterns of racial disparities in the adjudication of student misconduct and provide suggestive evidence of a mechanism that can explain their origins.

While school discipline can mitigate externalities associated with disruptive behavior (Carrell and Hoekstra 2010; Kinsler 2013; Pope and Zuo 2020), disciplinary interventions impose significant costs on disciplined students. Namely, 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 (Steinberg and Lacoe 2018; Anderson et al. 2019; Craig and Martin 2019; Bacher-Hicks et al. 2019; Sorensen et al. 2022). 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 (Kinsler 2011; Anderson and Ritter 2017; Welsh and Little 2018; Ritter and Anderson 2018; Gopalan and Nelson 2019; Barrett, McEachin, Mills and Valant 2021; Shi and Zhu 2022; Liu, Hayes and Gershenson 2022). 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

[1] For example, individual expectations about comparative advantage have been shown to influence consequential decisions about major choice in university settings (Arcidiacono et al. 2012; Card and Payne 2017).

(Kinsler 2011; Anderson and Ritter 2017; Ritter and Anderson 2018; Gopalan and Nelson 2019; Barrett et al. 2021). Yet gaps typically remain after conditioning on student characteristics and school fixed effects (Beck and Muschkin 2012; Gopalan and Nelson 2019; Anderson and Ritter 2020; Barrett et al. 2021; Shi and Zhu 2022; Liu et al. 2022), which leaves open the possibility that school officials treat students of color less favorably than white students who engage in similar behaviors. Even so, within-school comparisons may nevertheless fail to isolate the average response of school officials to the race of their students. Within-school disparities are consistent with differential treatment on the basis of race, but they are also consistent with systematic but unobserved differences in student behavior.

To adjust for unobserved differences in student behavior, several 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 among all incident types. 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. (2022) leverages incident identifiers in administrative data on student referrals, but from a large California school district. All three studies detect 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. 2022 and Shi and Zhu 2022). Together, these studies suggest that school administrators exhibit biases that disfavor 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.[2] Within-school disparities are typically larger in magnitude (e.g., Gopalan and Nelson 2019, Anderson and Ritter 2020, Barrett et al. 2021, Shi and Zhu 2022).

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

That being said, within-incident race differentials are identified from 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, when an administrator adjudicates students of color and white students "side-by-side," one would expect racial differences to be more salient, which may induce 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 and 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 would move them toward equal treatment in multi-race incidents. 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 et al. 1999; Hsee et al. 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. Regardless of the particular mechanism, however, 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. By relying on within-incident variation to identify race-based differentials in the adjudication of misconduct, we fear that within-incident comparisons identify the effect of race in an environment that is neither representative of the choices or incentives of administrators nor representative of the experiences of students of color.

We therefore seek to contextualize the adjudication of multi-race incidents by comparing

adjudication outcomes across joint incidents that are similar but differ in their racial composition. In multi-race fights, we find no race-based differences in outcomes. However, the magnitude of the difference in punishments across same-race fights within schools is often large—fights that involve only students of color elicit harsher punishments than those involving only white students. In high schools, for example, students of color in fights involving *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, in the same environment, 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, 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.

A causal interpretation of these findings suggests that the salience of differences in race within an incident moves administrators toward the equal treatment of students on the basis of race.[3] By the same token, the large disparities across same-race fights suggest that within-incident comparisons can severely understate the extent of differential treatment.

We proceed in Section 2 by describing the empirical environment and data we use to test for race-based differentials in punishment. We then present estimates of racial disparities in expulsions and suspensions in Section 3. Therein, we progress from raw race-based gaps in adjudication outcomes to within-incident differences. In Section 4 we incorporate same-race incidents to better understand the origins of the racial disparities we identify in Section 3. Finally, in Section 5, we discuss implications of our findings.

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 Superinten-

[3] We note that the results of the supplemental student fixed effects analysis in Shi and Zhu (2022) are not necessarily at odds with our findings. In that analysis, the authors consider how *relative* differences in adjudication outcomes vary across the racial composition of incidents that a student participates in. They find that Black students receive harsher punishments *relative to the other student in an incident* when the other student is of a different racial background. By measuring outcomes relative to others in the same incident, rather than levels, this analysis leaves open the possibility that the absolute severity of punishments imposed on students of color decreases in multi-race incidents.

dent of Public Instruction between 2014–15 and 2017–18.[4] Similar to Shi and Zhu (2022) and Liu et al. (2022), our administrative data include incident identifiers and infractions that did not result in students receiving exclusionary discipline. We complement Barrett et al. (2021), Shi and Zhu (2022), and Liu et al. (2022) by identifying disparities in a setting in which racial resentment is persistently lower (Smith et al. 2020).[5] 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 capturing race categorically, generally, 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 the 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.[6] That said, the qualitative conclusions from a more granular analysis of

[4] 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.

[5] This is also consistent with the data collected by Project Implicit, which suggests that implicit racial attitudes 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].

[6] Specifically, we define students of color as those who identify as (i) solely Black, (ii) Hispanic (of any race),

specific racial disparities (i.e., Black-white, Hispanic-white, and gaps between the remaining students of color and white students) are unchanged, and are similar to those documented in Section 3 and Section 4, but with less precision.

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.[7] 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 et al. 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

(iii) solely Asian, (iv) solely Pacific Islander, (v) solely Native American, or (vi) two or more races.

[7] 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.

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.[8]

There are a total of 66,355 fighting infractions among boys in our data. While schools are required to use the same incident identifier for incidents that involve multiple students, 33 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.[9] 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 discriminatory 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 16,279 infractions from 7,641 multi-student incidents that implicate at least two

[8] 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.

[9] 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.

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 findings are not sensitive to these restrictions.

2.2 Outcomes

We consider three margins of formal exclusionary discipline as outcomes for each infraction: whether the student is expelled, whether the student receives any suspension or expulsion, and the length of suspension conditional on all students being suspended within an incident.[10] In Table 1 we provide average disciplinary outcomes by school level (i.e., elementary, middle, high) 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. Fewer than one percent of fighting infractions in high school result in expulsion (the most severe punishment we observe in the data) and expulsions for fighting are rarer still in elementary and middle schools—too rare to build reasonable inference from. On other margins, punishments vary significantly across school levels—the rate of formal exclusionary discipline (suspension or expulsion) doubles between elementary and middle school, and the average suspension is over one day longer in high schools than in middle schools.[11] Within each grade span, average disciplinary outcomes are similar across samples, though exclusionary discipline rates are somewhat higher in the multi-student sample than in the larger samples, and suspensions are somewhat shorter.

2.3 Student characteristics

We observe a total of 41,520 students in the full sample and 12,855 students in the multi-student sample. Roughly 41 percent of students in the multi-student sample are white (non-Hispanic), 27 percent are Latino (Hispanic origin of any race), 16 percent are Black, 9 percent report more

[10] We exclude suspension lengths longer than 20 school days (approximately one calendar month) to limit the influence of rare long-term suspensions (over 99 percent of suspensions are shorter than 20 days). However, the inferences we make are not sensitive to perturbations of this 20-day cutoff.

[11] 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."

than one race, 3 percent are Asian, 2 percent are Pacific Islander, and 2 percent are Native American.

We derive controls for socioeconomic status, disability, and past achievement from an extended panel that dates back to 2009–10. We measure socioeconomic status using persistent eligibility for free or reduced-price meals. While there are significant racial disparities in socioeconomic status within each grade level, the vast majority of infractions in our sample implicate students from low-income households—this is true for white students and students of color alike. Using up to nine years of data, we determine whether a student is (i) always eligible, (ii) never eligible, or (iii) sometimes 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 and Dynarski 2017). To measure special-education status, which is an important predictor of punishment, we derive 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. We control for observed English Language Arts (ELA) and math achievement levels from the most-recent grade tested. 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 students of color. 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.[12]

Using an additional year of infraction data, we also control for each student's infraction history, measured as the number of infractions from the previous school year. Students who select into fights typically have an infraction from the previous school year, and students of color tend to have more past infractions than their white peers.[13]

[12] For this reason, we include students with test scores that are unobservable to both the econometrician and school administrators in our analysis (e.g., elementary students who have not yet been tested, as tests are not available until the third grade). We allow for any level differences for those without test scores with an indicator variable, though results are robust to their exclusion.

[13] The results we report in Section 3 and Section 4 are not sensitive to controlling specifically for the number of fighting infractions from the previous school year.

3 Within-incident comparisons

Given the potential for differential selection into incidents (by students, for example) and the potential for differential adjudication of incidents (by different 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 (i.e., 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 actors are implicated as part of the mechanism that produces differential outcomes.

Below, we consider expulsion and the extensive and intensive margins of suspension, 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 one may be more inclined to interpret estimates as causal. We will then re-direct our efforts toward identifying a mechanism that can explain the advent of race-based differentials in punishment.

3.1 Expulsions

In Figure 1 we begin by reporting unconditioned differences in the adjudication of student misconduct, and then progressively restrict the identifying variation that contributes to the estimated difference. The left-most estimate in Panel A is the raw difference in expulsion rates between white students and students of color in high school. Conditional on receiving an infraction, expulsion rates for students of color are 0.66 percentage points, or 213.2 percent, higher ($p < 0.001$) than those for white students. Relative to the sample standard deviation (σ) of expulsion in the estimation sample, this difference corresponds to an effect size on the order of 0.08σ.

In Column (2) we control for student attributes (e.g., grade, eligibility for free or reduced-price school meals, past achievement, and proxies for the receipt of special education services), past infractions, and school-by-year fixed effects, absorbing any variation in punishment across schools into the error term for the sample of all fighting infractions. In contrast to Kinsler (2011), who estimates a similar specification for suspensions using data from North Carolina, this fails to decrease the variation in expulsion that is attributable to race, and within-school variation in expulsions are still suggestive of significant gaps in the adjudication of infractions for students of color compared to white students. The introduction of school-by-year fixed effects does attenuate race differentials for suspension outcomes, considered further below, but significant gaps remain nonetheless.

In Column (3) 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 restrict identifying variation to within-incident variation. As in the full sample, the unconditioned gap in these schools implies that students of color are significantly more likely than white students to be expelled for fighting. This difference is larger in magnitude than the point estimate in Column (1), though the confidence intervals do overlap. Likewise, the addition of controls and school-by-year fixed effects in Column (4) has little impact on the magnitude of the estimated race differential.

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 past infractions.

We begin to approach something that may defensibly justify a causal interpretation in the Column (8) of Figure 1, where we control for school heterogeneity with the inclusion of school-by-year fixed effects. However, it is in columns (11) through (13) that we absorb any unobserved heterogeneity that is specific to incidents—this is where we are most confident in having retrieved estimates that warrant a causal interpretation. We execute these within-incident comparisons by estimating models of the form

$$\mathbb{1}(\text{Expelled} = 1)_{ikst} = \beta\ \text{SoC}_i + X_{ij}^{'}\Theta + \lambda_k + \nu_{ikst}\ , \tag{1}$$

where $\mathbb{1}(\text{Expelled} = 1)_{ikst}$ captures the expulsion of student i associated with their involvement in incident k in school s during year t.[14] Incident fixed effects (λ_k) capture unobserved heterogeneity in incidents (nested within schools). Student controls ($X_i^{'}$) adjust for level differences that arise from within-incident variation in student attributes (i.e., grade, 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 past infractions (from the previous school year). Our parameter of interest (β) absorbs the average difference in expulsion rates for students of color ($\text{SoC}_i = 1$) relative to white students ($\text{SoC}_i = 0$). The error term (ν_{ikst}) captures any remaining variation, and we allow for clustering at the school-by-year 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 $\hat{\beta}$—we assume that students of color are not differentially culpable, conditional on the full set of controls and incident fixed effects. 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 discriminatory adjudication. That said, across the fighting infractions we consider, we are less concerned that differential selection into incidents explains our results—the severity of these behaviors presents teachers with few opportunities to exercise discretion in deciding whether to refer students to the principal's office for discipline.

We find no statistically significant difference in the probability of expulsion for students of color relative to white students in the same incident. In the preferred specification, the probability of expulsion is 0.48 percentage points higher (227%, 0.08σ), on average, for students of color, but the difference is indistinguishable from zero at conventional significance levels ($p = 0.339$). Though the difference in probability is statistically insignificant, note that we

[14] No student (i) has multiple infractions (j) within the same incident (k).

cannot rule out meaningful effect sizes at the upper bound of the 95-percent confidence interval.

3.2 Suspensions

In panels B and C of Figure 1 we repeat a similar exercise for suspensions—in Panel B we model the extensive margin, and in Panel C we estimate models of suspension length conditional on suspension.[15] As expected, suspensions vary systematically with race—unconditioned, students of color experience rates of formal exclusionary discipline that are 5.59 percentage points higher (15.1%, 0.11σ, $p < 0.001$) in elementary schools, 3.97 percentage points higher (4.9%, 0.11σ, $p < 0.001$) in middle schools, and 1.52 percentage points higher (1.7%, 0.05σ, $p = 0.023$) in high schools. Racial disparities are also large on the intensive margin of suspension—conditional on being suspended for fighting, students of color receive suspensions that are 0.08 days longer (5.8%, 0.06σ, $p = 0.013$) in elementary schools, 0.31 days longer (14.8%, 0.18σ, $p < 0.001$) in middle schools, and 0.51 days longer (15.9%, 0.21σ, $p < 0.001$) in high schools. However, when we restrict the identifying variation to that existing within incidents we find that students of color are neither suspended at higher rates than white students implicated in the same incident, nor suspended for any longer, conditional on being suspended. Within-incident variation in student race does not support the claim that there are significant differences in suspensions experienced by students of color, *on average*.

In fact, 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 across elementary, middle, and high schools. Although some estimates have relatively wide confidence intervals—namely those of expulsion and suspension length 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

[15] For the analysis in Panel B we model "suspended or expelled" together as there are two margins around which we anticipate student selection. Namely, there are students who are at the margin of being suspended, and (likely different) students who are at the margin of being expelled—defined this way, exit is necessarily toward lesser consequences. For the analysis in Panel C we restrict the sample to incidents that result in suspensions for all students.

fight, then differential treatment within incidents explains very little of the aggregate racial disparities.

4 Where do gaps arise, then?

One explanation for the absence of significant gaps in punishment within multi-race fights is that within-incident variation in race offers a degree of salience that enables the equal treatment of students of color. For example, it could be easier for administrators to suppress implicit biases within incidents. Alternatively, it could be that explicit biases are more costly to act on within incidents, where one cannot appeal to incident heterogeneity (e.g., "It was a really bad fight") as a justification for harsher punishment.

That punishment gaps attenuate when we identify off of within-incident variation is also consistent with race-based differences in parents' inclinations to advocate for their children, or for their advocacy to exert varying degrees of influence on punishments. For example, if advocacy varies more across race than within, we should expect advocacy-driven variation in outcomes to be partially absorbed by the incident fixed effect. However, for a differential-advocacy story to explain the variation we see in the data, it would need to be the case that administrators respond to the advocacy given to a white student *and* extend it to others involved in the same fight, regardless of race. In that way, administrators still appear better able to maintain equality norms within fights than they do across fights.

By absorbing the unobserved heterogeneity associated with specific incidents, the foregoing analysis 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 Figure 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.[16] Specifically, we estimate models of the form

[16] Multi-race fights make up 34.8 percent of multi-student fights, while 42.5 percent implicate only students of color—the remaining 22.7 percent implicate only white students.

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

where Punishment_{ikst} is the disciplinary intervention assigned to student i associated with their involvement in incident k in school s during year t. As before, we control for student attributes and past infractions, and identify racial gaps across same-race fights (β) with the conditional variation that exists within the same school during the same school year. As selection into multi-race fights may differ, we absorb any level effect associated with multi-race fights in τ.[17] 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.

A causal interpretation of the within-incident differences in Figure 1 required the assumption that students of color are not differentially culpable, conditional on controls and incident fixed effects. In Figure 2, however, to interpret $\hat{\phi}$ as causal we must also be willing to assume that there is no differential selection into multi-race fights—it cannot be that it is the less-culpable students of color who are selecting into fights with white students. In other words, to explain away the variation we observe in the data (conditioning on school-by-year fixed effects, student characteristics, and past infractions) one must simultaneously believe 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.

4.1 Results

In Figure 2 we plot two coefficient estimates from each model. The first is the estimated difference in outcomes for students of color, identified off of same-race fights (i.e., the average

[17] Point estimates of τ are generally positive, capturing that multi-race fights tend to be punished more heavily than same-race fights. In high school expulsions and suspension length, and in elementary school suspension length, $\hat{\tau}$ is small but statistically significant, though in all models $\hat{\tau}$ is smaller in magnitude than both $\hat{\beta}$ and $\hat{\phi}$.

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 outcomes for students of color in multi-race fights (i.e., the average difference in outcomes experienced by students of color in fights that implicated a white student).

In Panel A we consider expulsions among high school students. Our preferred specifications identify that (i) students of color *in fights that only implicate students of color* experience significantly higher rates of expulsion (2.24 percentage points, 0.27σ, $p = 0.006$) and (ii) the increase in expulsion rates for students of color is offset when there is a white student implicated in the same fight (-1.75 percentage points, -0.21σ, $p = 0.022$). 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 (0.49 percentage points, 0.06σ, $p = 0.324$), which is consistent with the presence of a white student *fully* offsetting the average difference in expulsion rates. The expulsion gap persists when we restrict the sample to fights that involve exactly two students (1.26 percentage points, 0.2σ, $p = 0.041$), as does the offsetting difference for being implicated with a white student (-1.38 percentage points, -0.22σ, $p = 0.047$). As in the multi-student sample, the sum of the coefficients indicates that students of color are no more likely to be expelled when they are implicated with white students (-0.12 percentage points, -0.02σ, $p = 0.798$). In either sample, "impact" estimates—the percentage changes over the mean of the reference group—are undefined, as no white student in an all-white incident is expelled for fighting.

In Panel B of Figure 2 we consider suspension disparities for each grade level. Again, we find that students of color are only more likely to experience exclusionary punishment when they are implicated with only other students of color. This is most evident in middle schools, where students of color are 4.6 percentage points more likely to receive exclusionary punishment when they are implicated with only students of color (5.3%, 0.2σ, $p = 0.003$), but no more likely when they are implicated with white students (-0.04 percentage points, 0%, 0σ, $p = 0.96$). The same pattern is also evident in elementary schools, where students of color experience a larger gap when they are implicated with only students of color (3.47 percentage points, 7.7%, 0.12σ, $p = 0.214$) than when they are implicated with white students

(1.48 percentage points, 3.3%, 0.05σ, $p = 0.128$), though precision falls short of conventional significance levels. In high schools, however, we do not observe significant differences for students of color when they are implicated with only students of color (−0.86 percentage points, −0.9%, −0.05σ, $p = 0.593$) or when they are implicated with white students (−0.1 percentage points, −0.1%, −0.01σ, $p = 0.894$). Across all grade levels, inferences are robust to restricting the sample to fights that involve exactly two students.

In Panel C we consider the intensive margin of suspension, where we find similar patterns, and with enough precision to suggest that the patterns we identify are a significant part of the data-generating process. Relative to white students in same-race fights, suspensions for students of color in fights that implicate only other students of color are, on average, 0.33 days longer in elementary school (26.9%, 0.57σ, $p = 0.002$), 0.14 days longer in middle school (7.3%, 0.11σ, $p = 0.048$), and 0.68 days longer in high school (22.6%, 0.41σ, $p < 0.001$). However, when implicated with white students, students of color receive suspensions that are no longer than those for white students in all-white fights—this is true in elementary school (0.06 days, 4.3%, 0.09σ, $p = 0.281$), in middle school (−0.01 days, −0.6%, −0.01σ, $p = 0.802$), and in high school (0.04 days, 1.6%, 0.03σ, $p = 0.607$). As in Panel B, this pattern is robust to restricting the sample to fights that involve exactly two students.

The offsetting differences in Figure 2 suggest that the within-school disparities in Figure 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, than white students implicated for fighting in the same school during the same school year. When implicated in fights without white students, however, students of color receive systematically harsher punishments than those imposed on white students.[18]

Depending on the grade level and margin of punishment, the magnitude of the difference in punishment across 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

[18] Recall that the vast majority of infractions in our sample are from low-income students—this is true for both students of color and white students. The relative dearth of students from higher-socioeconomic-status backgrounds increases our confidence that the patterns we document in Figure 2 reflect the salience of race rather than the salience of socioeconomic status. Estimates of β and ϕ are qualitatively similar when we stratify each sample by socioeconomic status and run them separately, though we lose precision in the sub-samples with higher socioeconomic status.

estimate across same-race fights—especially those on the intensive margin of suspension—are larger in magnitude than within-school gaps documented elsewhere in the literature (e.g., Kinsler 2011; Barrett et al. 2021; Anderson and Ritter 2020; Shi and Zhu 2022).

If students of color and white students in same-race fights are equally culpable (after conditioning on school-by-year fixed effects and the full set of controls), then the empirical regularities 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 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 et al. 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 (Lacoe and Steinberg 2018; Steinberg and Lacoe 2018; Craig and Martin 2019; Pope and Zuo 2020), while others have experimented with less punitive disciplinary interventions, such as restorative justice (Glenn et al. 2020). However, the ultimate success of any disciplinary reform depends, in part, on the ability of school officials to enforce policies without partiality. With evidence that educators hold biases that disfavor students of color (Chin et al. 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.

In our analysis, we inquire into the source of racial disparities in the adjudication of student misconduct. In incidents that implicate at least two boys for fighting, we find that students of color receive harsher punishments, on average. Across same-race fights in the

same school during the same school year, students of color are more likely to be suspended or expelled than white students, and tend to receive longer suspensions conditional on being suspended. However, within multi-race fights, the punishments imposed on students of color are statistically indistinguishable from those imposed on white students. Moreover, we document a pattern, evident across grade levels and robust to a variety of alternative specifications, in which the presence of a white student *fully* offsets within-school punishment differentials for students of color.

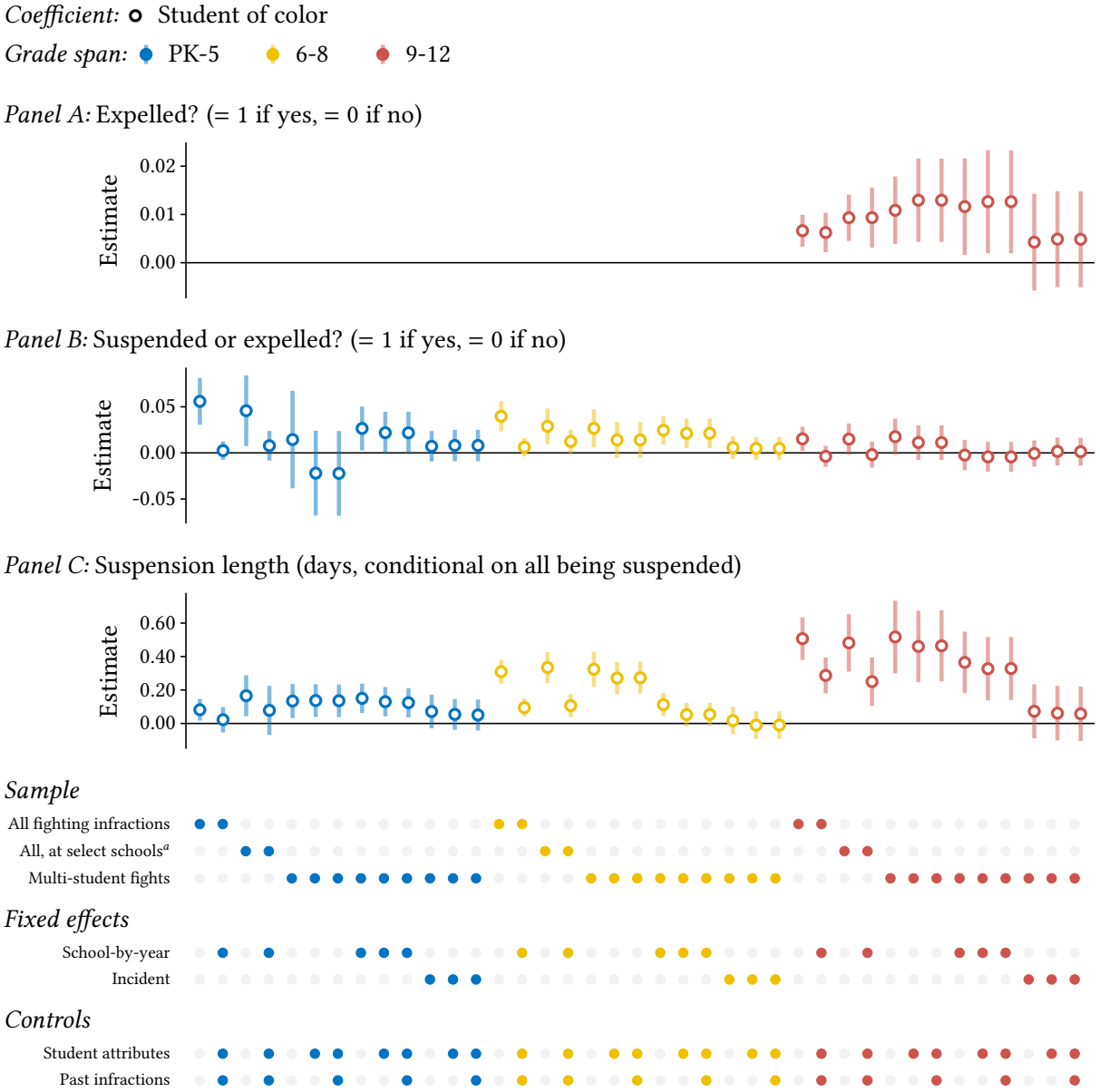
We find encouragement insofar as the data-generating process supports that biases are correctable where race and equality norms are more salient. That being said, our results imply that purging all within-incident disparities in punishment would do little to close the gap in disciplinary outcomes between 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 discrimination—relying on within-incident comparisons to monitor unequal treatment would falsely signal an equality in outcomes, understating the extent of differential treatment.

Table 1: Summary statistics

	Grades PK–5		Grades 6–8		Grades 9–12	
	μ (1)	σ (2)	μ (3)	σ (4)	μ (5)	σ (6)
Panel A: All fighting infractions						
Expelled? (= 1 if yes, = 0 if no)	0.00	0.01	0.00	0.04	0.01	0.08
Observations	29,523		24,999		11,832	
Incidents	27,025		20,979		9,799	
Suspended or expelled? (= 1 if yes, = 0 if no)	0.40	0.49	0.83	0.37	0.91	0.28
Observations	29,523		24,999		11,832	
Incidents	27,025		20,979		9,799	
Suspension length (days)	1.45	1.27	2.26	1.72	3.48	2.39
Observations	11,424		20,264		10,325	
Incidents	10,430		16,839		8,562	
Panel B: All fighting infractions from schools that report at least one multi-student fight						
Expelled? (= 1 if yes, = 0 if no)	0.00	0.00	0.00	0.04	0.01	0.09
Observations	12,018		14,599		7,026	
Incidents	9,533		10,582		5,002	
Suspended or expelled? (= 1 if yes, = 0 if no)	0.40	0.49	0.86	0.35	0.92	0.28
Observations	12,018		14,599		7,026	
Incidents	9,533		10,582		5,002	
Suspension length (days)	1.54	1.39	2.30	1.74	3.54	2.41
Observations	4,572		12,132		6,125	
Incidents	3,588		8,709		4,367	
Panel C: Multi-student fights						
Expelled? (= 1 if yes, = 0 if no)	0.00	0.00	0.00	0.04	0.01	0.09
Observations	4,556		7,587		3,560	
Incidents	2,141		3,703		1,666	
Suspended or expelled? (= 1 if yes, = 0 if no)	0.44	0.50	0.89	0.31	0.94	0.24
Observations	4,556		7,587		3,560	
Incidents	2,141		3,703		1,666	
Suspension length (days)	1.35	0.83	2.16	1.51	3.47	2.22
Observations	1,868		6,495		3,159	
Incidents	905		3,181		1,501	

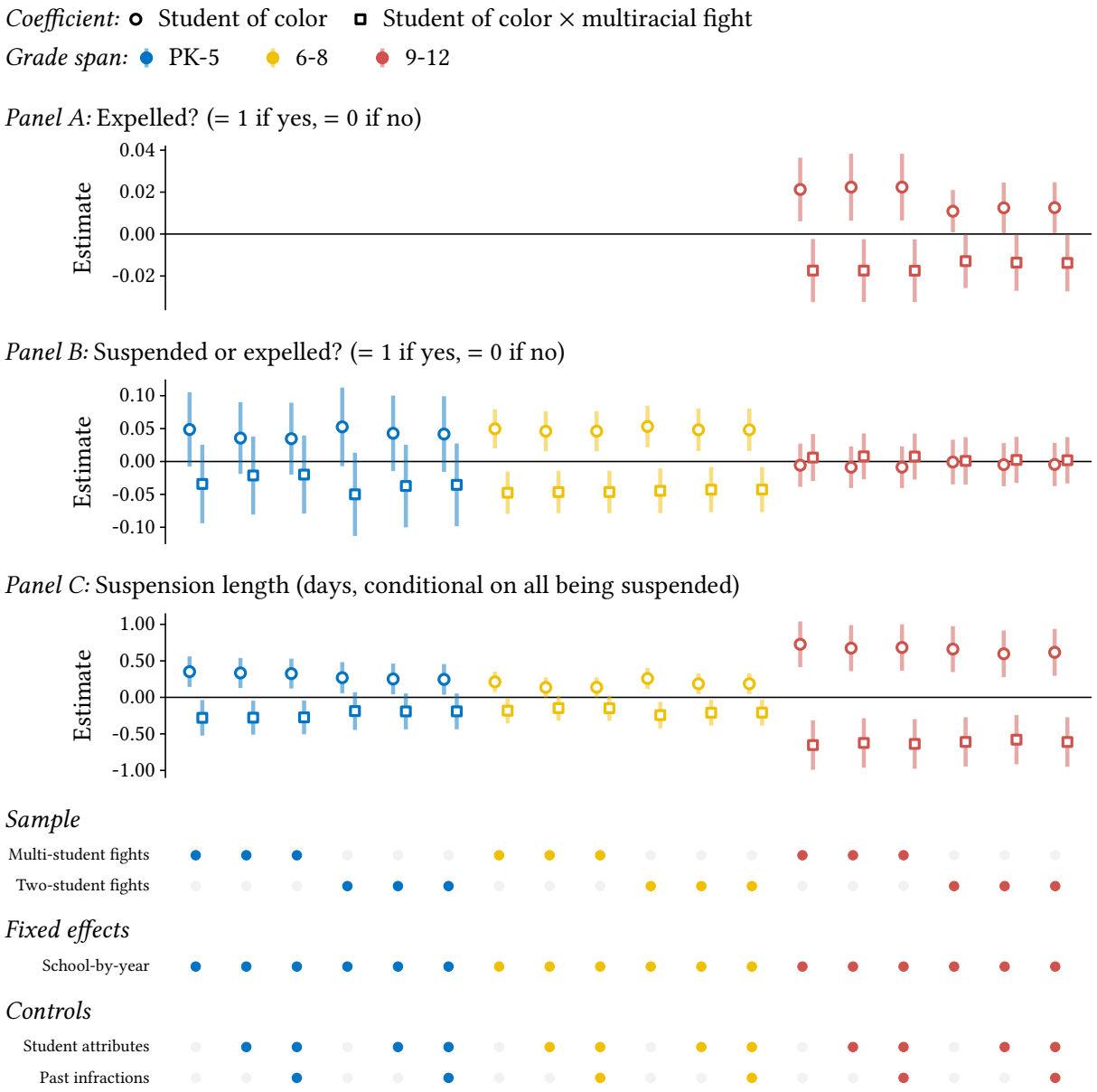
Notes: Sample means (μ) and standard deviations (σ) of punishment outcomes considered in Section 3 and Section 4. The alternative to an expulsion or a suspension is either "no intervention" or "other intervention." Suspension lengths are conditional on all students being suspended within each incident. The sample in Panel A consists of boys' infractions for "fighting without major injury" between 2014–15 and 2018–18. 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.

Figure 1: Punishment disparities in school fights



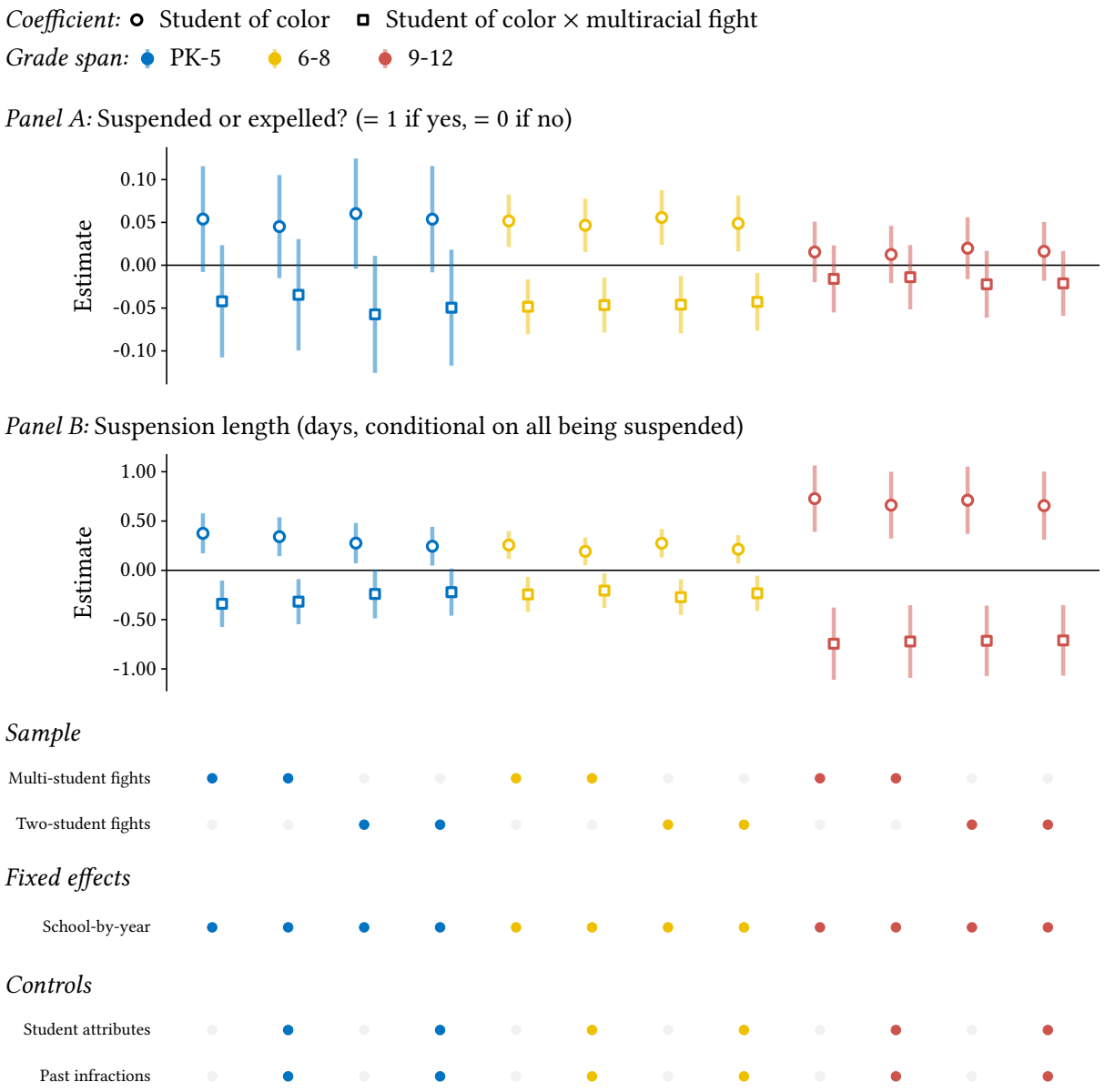
Notes: Open circles show OLS estimates of racial punishment gaps. Each estimate is from a different regression. The leftmost estimate in each grade span describes a raw punishment gap, and the rightmost estimate describes a within-incident punishment gap from the fully specified model (e.g., see Equation 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-by-year level.
[a]All fighting infractions from schools that report at least one multi-student fight.

Figure 2: Punishment disparities across school fights by racial composition



Notes: Open circles and squares show OLS estimates of coefficients from Equation 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-by-year level.

Figure 3: Punishment disparities across school fights by racial composition (Latino, Black, and two or more races only, and white only)



Notes: Open circles and squares show OLS estimates of coefficients from Equation 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-by-year level.

References

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

_ and _, "Do school discipline policies treat students fairly? Evidence from Arkansas," *Educational Policy*, 2020, *34* (5), 707–734.

_ , _ , and Gema Zamarro, "Understanding a vicious cycle: The relationship between student discipline and student academic outcomes," *Educational Researcher*, 2019, *48* (5), 251–262.

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

Bacher-Hicks, Andrew, Stephen B Billings, and David J Deming, "The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime," Working Paper 26257, National Bureau of Economic Research September 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*, 2021, *56* (3), 711–748.

Bazerman, Max H, Don A Moore, Ann E Tenbrunsel, Kimberly A Wade-Benzoni, and Sally Blount, "Explaining how preferences change across joint versus separate evaluation," *Journal of Economic Behavior & Organization*, 1999, *39* (1), 41–58.

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

Card, David and A Abigail Payne, "High school choices and the gender gap in STEM," *Economic Inquiry*, 2017.

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

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

Chin, Mark J, David M Quinn, Tasminda K Dhaliwal, and Virginia S Lovison, "Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes," *Educational Researcher*, 2020, *49* (8), 566–578.

Craig, Ashley C and David C Martin, "Discipline Reform, School Culture, and Student Achievement," 2019.

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

Fudenberg, Drew and Jean Tirole, "A "Signal-Jamming" Theory of Predation," *The RAND Journal of Economics*, 1986, *617* (3), 366–376.

Glenn, Beth, Nathan Barrett, and Estilla Lightfoot, "The Effects and Implementation of Restorative Practices for Discipline in New Orleans Schools," Technical Report, Education Research Alliance for New Orleans 2020.

Gopalan, Maithreyi and Ashlyn Aiko Nelson, "Understanding the racial discipline gap in schools," *AERA Open*, 2019, *5* (2).

Holmström, Bengt, "Managerial Incentive Problems: A Dynamic Perspective," *The Review of Economic Studies*, 1999, *66* (1), 169–182.

Hsee, Chrisopher K, George F Loewenstein, Sally Blount, and Max H Bazerman, "Preference reversals between joint and separate evaluations of options: A review and theoretical analysis," *Psychological Bulletin*, 1999, *125* (5), 576.

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

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

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

Liu, Jing, Michael S Hayes, and Seth Gershenson, "JUE Insight: From Referrals to Suspensions: New Evidence on Racial Disparities in Exclusionary Discipline," *Journal of Urban Economics*, 2022.

Losen, Daniel J, Cheri L Hodson, Michael A Keith II, Katrina Morrison, and Shakti Belway, *Are we closing the school discipline gap?*, Los Angeles: The Center for Civil Rights Remedies, University of California, 2015.

Michelmore, Katherine and Susan Dynarski, "The gap within the gap: Using longitudinal data to understand income differences in educational outcomes," *AERA Open*, 2017, *3* (1).

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

Reykdal, Chris, Katie Weaver-Randall, and Lisa Ireland, "Comprehensive Education Data and Research System (CEDARS) appendix manual, version 10.2," Technical Report, Washington State Office of Superintendent of Public Instruction January 2018.

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

Shi, Ying and Maria Zhu, "Equal time for equal crime? Racial bias in school discipline," *Economics of Education Review*, 2022, *88*.

Smith, Candis Watts, Rebecca J Kreitzer, and Feiya Suo, "The dynamics of racial resentment across the 50 US states," *Perspectives on Politics*, 2020, *18* (2), 527–538.

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*, 2022, *17* (2), 255–284.

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

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