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

Kyle Raze *　　Glen R. Waddell †

September 2021

Abstract

Racial gaps in schools' disciplinary actions are well documented—for similar behaviors, students of color are more likely to be disciplined and discipline tends to be harsher. Using incident-level data, we identify differential treatment across the racial composition of incidents. While students of color receive harsher punishments on average, we show that this differential is driven by incidents without white students. Consistent with administrators correcting biases when cues for equal treatment are more salient, multiracial incidents evidence no differentials—when a white student is implicated in the same incident, punishments imposed on students of color are indistinguishable from those of white students.

Keywords: student behavior, school discipline, racial discrimination

JEL Classification: I24, J71

*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 so many outcomes, any degree of hysteresis in the production of race-based bias 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 for us to consider. In fact, we should worry about race-based gaps in exactly this environment—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 practice, the set of disciplinary actions commonly available to school administrators are also inseparable from interruptions to the direct inputs into the production of human capital—exclusionary discipline both decreases instructional time and interrupts the continuity of instruction. Such interruptions pose meaningful consequences, then—exclusionary discipline can hinder academic performance, decrease educational attainment, and increase the likelihood of arrest and incarceration (Steinberg and Lacoe 2018; Craig and Martin 2019; Anderson et al. 2019; Bacher-Hicks et al. 2019). Understanding *where gaps arise* is important, as unexplained gaps are both morally objectionable and easily tied to significant and long-lasting economic inefficiencies in the production of human capital.

In terms of existing research, there is the established belief that a large portion of the *average* discipline gap arises across schools, with Black students being over-represented in schools and school districts with higher rates of exclusionary discipline (Anderson and Ritter 2017; Ritter and Anderson 2018; Gopalan and Nelson 2019; Barrett et al. 2019). However, gaps often remain after conditioning on student characteristics and school fixed effects (Beck and Muschkin 2012; Gopalan and Nelson 2019; Barrett et al. 2019). Identifiers that allow for the absorption of unobserved incident heterogeneity are usually unavailable, so the inclusion of "school fixed effects" is typically the extent to which researchers have been able to restrict identifying variation in disciplinary outcomes. And, of course, identifying any potential differential in how students of color are treated, even when implicated in the *same* incident,

[1] Though in a university setting, expectations have been shown to influence consequential decisions about major choice, for example (Arcidiacono et al. 2012; Card and Payne 2017).

will be informative.

In our case, our fully specified models reveal no evidence of significant race-based gaps—this is true across all grade levels and exclusionary outcomes. However, while one might be inclined to interpret models with incident fixed effects as providing "better" identification, by controlling for incident-specific heterogeneity, it is just as important to recognize this as *different* identification. To that end, while we will witness the collapse of race-associated gaps in disciplinary outcomes as we absorb more of the identifying variation into the error structue, the takeaway we will emphasize is that *existing gaps in the adjudication of infractions are not coming from differential treatment within multiracial incidents.* Out of that analysis, then, it will remain important to ascertain where raw gaps do emerge—if restricting identifying variation to that *within* incidents leaves no gap in punishments, we should not conclude that there is no gap but should continue to inquire into what drives the aggregate disparity.

To do this, we construct what we believe to be the cleanest test available given existing data. From the universe of recorded infractions in the state of Washington we restrict our attention to fights among boys, which are the most-common type of multi-student incident in the data, but also confer several additional benefits.[2] For example, fights are included in mandatory federal reporting and, as such, feature a clear definition of what constitutes misconduct (Reykdal et al. 2018). Consider the comparison of "fighting without major injury" and "disruptive conduct"—state guidance defines fighting without major injury as "mutual participation in an incident involving physical violence" where no student or school official requires professional medical attention. The state also provides examples of disqualifying injuries (e.g., "stab or bullet wounds, concussions, fractured or broken bones, or cuts requiring stitches"), and directs school officials to *exclude* "verbal confrontations, tussles, or other minor confrontations." (It is in examples like these that one would anticipate non-random selection into infractions. By restricting the sample to fights we avoid this source of potential confoundedness.) In contrast, the state's definition of disruptive conduct—behavior "that materially and substantially interferes with the educational process"—leaves much-more room for subjectivity, and for

[2] Specifically, we consider all infractions for "fighting without major injury." 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.

race-based selection into reported infractions.

Well-defined conditions and mandatory reporting help us imagine that selection into fights is less likely to depend on the subjective judgments of teachers than will selection into other infractions (Welsh and Little 2018), so our focus on fights tips toward limiting potential measurement error in the classification of incidents. Thus, 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. In terms of our framing of the external validity and context of our empirical test, restricting our analysis to fights also implies that we are more likely to have captured the set of relevant actors—while there will surely be some students who escape the eyes of teachers, the "jointness" of fights leaves us more confident that we will be identifying off of meaningful variation in the data-generating process.[3]

In the end, we bring new evidence to bear on the role of race in school discipline. This evidence is suggestive of concerning disparities in the experience of students of color but, at the same time, suggestive of mechanisms that promotes a richer understanding of their origins.[4] In particular, we will identify that (i) gaps do not originate in the unequal treatment of students in multiracial incidents, (ii) incidents that implicate only students of color tend to elicit harsher punishments than those that implicate only white students, and (iii) when implicated *with* white students, the treatment of students of color is indistinguishable from that of white students in all-white incidents.

In Section 2 we describe the data we use to test for race differentials in punishment, with our estimates of racial disparities in expulsions and suspensions in Section 3. Therein, we progress from raw race-based gaps in disciplinary outcomes through to within-incident differences. In Section 4 we incorporate same-race incidents to better understand the mechanism that generates the gaps in aggregate data. Finally, in Section 5 we discuss implications of the patterns we observe.

[3] 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 be missing race components to outcomes. Considering fights between students—fights being well-defined and subject to mandatory reporting—again mitigates such concerns.

[4] In doing so, we contribute to an extensive empirical literature on racial disparities in school discipline. See Welsh and Little (2018) for a comprehensive review of the literature.

2 Data

To document racial differences in the severity of sanctions for alleged misconduct, we consider all disciplinary infractions reported by Washington public schools to the Office of Superintendent of Public Instruction (OSPI) between 2014–15 and 2017–18.[5] Disciplinary infractions are the result of an authority figure (e.g., teacher, recess monitor, security officer) sending a student to a school administrator (e.g., the principal, vice principal) to receive discipline for alleged misconduct. Similar to infraction data available from other states, Washington's data provide a coded description of the alleged behavior and the type, length, and dates of punishment associated with each infraction.[6] Important to our ability to detect evidence of differential treatment is the availability of incident identifiers that allow us to identify when multiple students were involved together in an incident (i.e., the *same* identifier used for all students in a given incident).

While our data will facilitate the ability to identify incidents involving multiple students, 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

[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.

[6] The behavior codes describe 21 different types of misconduct. In order of frequency, the behavior codes are "disruptive conduct," "other behavior resulting in intervention," "failure to cooperate," "fighting without major injury," "violence without major injury," "marijuana," "bullying," "tobacco," "harassment, intimidation, or bullying," "possession of a weapon," "theft or possession of stolen property," "multiple minor accumulated incidents," "alcohol," "discriminatory harassment," "illicit drug (other than marijuana)," "sexual harassment," "destruction of property/vandalism," "sexually inappropriate conduct," "academic dishonesty/plagiarism," "violence with major injury," and "serious bodily injury."

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.[7] That said, the qualitative conclusions from a more-granular analysis of 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

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 are from schools that never report the same incident identifier for multiple students. Thus, our analysis will speak only to schools that follow the reporting guidelines.[8] Somewhat more troubling, potentially, is that in schools that do report matching incident identifiers, only 37 percent of fighting infractions have an incident identifier that matches that of another student in the infraction data. As a worst case, for example, one might imagine that white students systematically avoid fighting infractions, leaving an “excess” of students of color among the reported fights. (We note that 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.) However, we find no evidence of “missing” white students among fighting infractions without a matching identifier—in the full sample of boys’ fighting infractions, race does not predict the existence of a match within the same school and year. Yet, the safest inference going forward might be to interpret our estimated differentials as lower bounds of the effect of race on outcomes.

We derive our main findings from fighting infractions with *matching incident identifiers*. Identifying race differentials *within* incidents will be key to our preferred specifications in

[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] 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.

Section 3, and in Section 4 we leverage incident identifiers to compare punishments by the racial composition of fights. In total, we observe 16,279 infractions from 7,641 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 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.[9] 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.[10] 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.

[9] 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.

[10] 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."

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 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.[11]

Using an additional year of infraction data, we also control for each student's infraction

[11] 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.

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.[12]

3 Toward "better" identification

To identify any differential in the punishment of students by race, we would ideally vary the racial composition of incidents at random and measure school administrators' perceptions of incident severity and student culpability, likely expressed in punishment severity. However, 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, for example, 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 specification 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—a mechanism with clearer implications for policy.

[12] 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.

3.1 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 imagine the most confidence in having retrieved estimates that warrant a causal interpretation. This specification can be notated as

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

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.[13] 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

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

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 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.[14] 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

[14] 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.

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—other 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, on average, white students and students of color are equally culpable for their involvement in a fight then differential treatment *within incidents* explains very little of the aggregate racial disparities.

In the most closely related analysis, Barrett et al. (2019) constructs a sample of unique fights by restricting incidents in Louisiana public schools to fighting infraction that were on the *same day* in the *same school*, and for which *exactly two students were suspended*. Disciplinary gaps in Barrett et al. (2019) also attenuate as one approaches "better" identification. However, being limited to considering only the intensive margin of suspension exposes the inference in Barrett et al. (2019) to potential selection along the extensive margin of suspension—we find no significant gaps on the extensive margin, though, which should strengthen the takeaway, that in Louisiana public schools there are small within-incident racial disparities in suspension lengths in middle school and in high school. We find smaller estimates in middle school—we are inclined to interpret ours as "zero"—and smaller estimates in high school, though less precise.

4 Where do gaps arise, then?

One explanation for the absence of significant gaps in punishment within multiracial fights is that within-incident variation in race offers a degree of salience that enables the equal treatment of students of color—perhaps administrators can more easily suppress implicit biases within incidents, for example. 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 multiracial 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 multiracial and same-race fights.[15] Specifically, we estimate the following set of models:

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

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 multiracial fights may differ, we absorb any level effect associated with multiracial fights in τ.[16] 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

[15] Multiracial 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.

[16] Point estimates of τ are generally positive, capturing that multiracial 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 $\hat{\phi}$.

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 multiracial 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 together are somehow more deserving of punishment than white students who select into fights together 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 together.

4.1 Results

In Figure 2 we plot two coefficient estimates from each of these models. The first is the estimated difference in outcomes for students of color, identified off of same-race fights—average outcomes across fights that implicated only students of color compared to fights that implicated only white students. The second, then, is the estimated difference in outcomes for students of color in multiracial fights—the average difference in the outcome 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 multiracial 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 only with other students of color (i.e., with no white student). 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 than white students who are 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.[17]

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 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. 2019; Anderson and Ritter 2020).

If students of color and white students in same-race fights are similarly 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 that only involve students of color. The full characterization of the data-generating process—with within-incident variation coming from multiracial fights—strongly suggests that the presence of a white student *moves administrators toward equal treatment*, consistent with administrators correcting biases when race is more salient.

[17] 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.

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 rolled back strict "zero tolerance" discipline policies (which according to Curran (2016) may have had a disparate impact on students of color) in favor of 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) or promote less-punitive 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), whether those biases manifest in disparate treatment remains an important question for researchers.

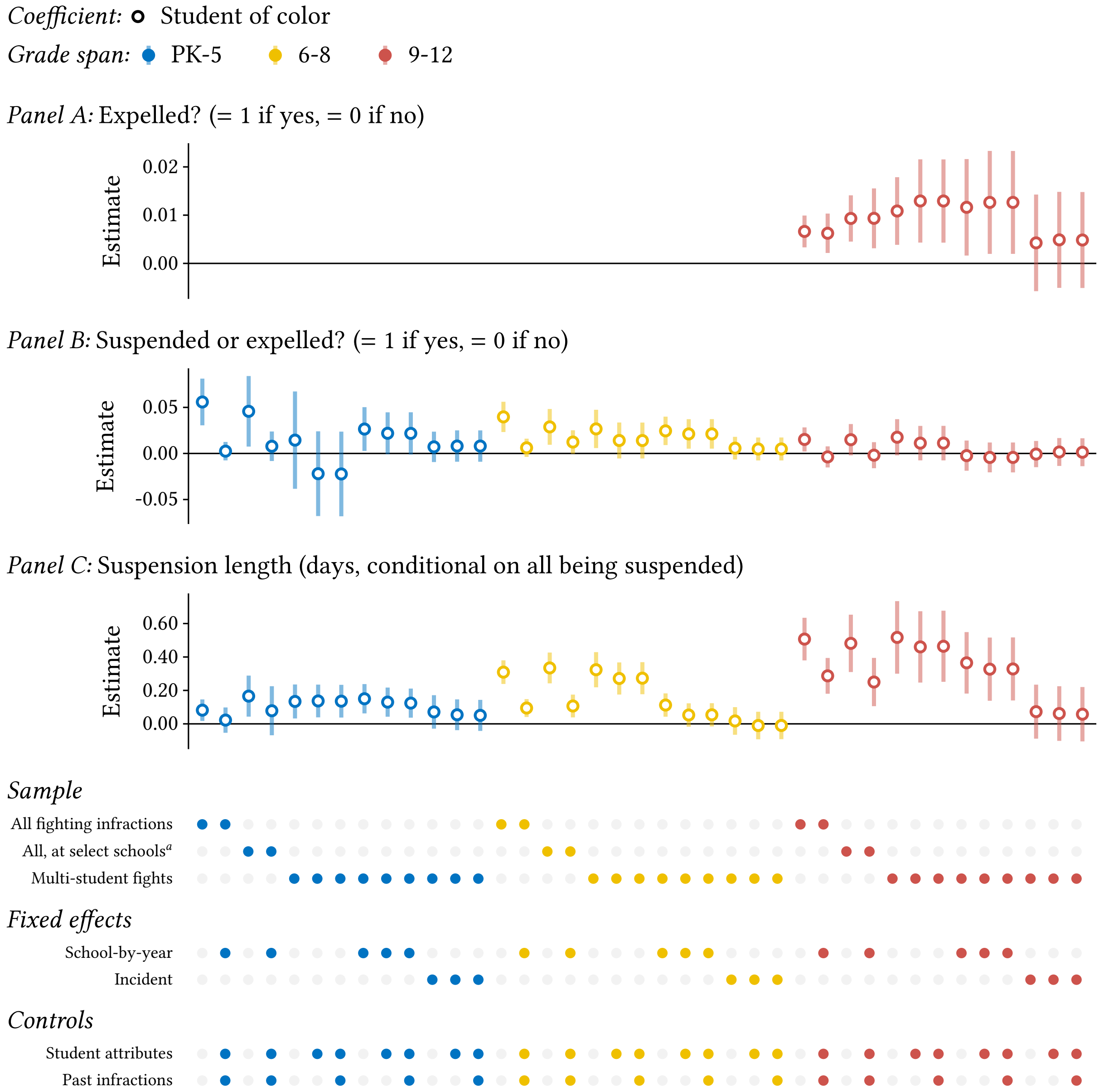
We address the question of differential treatment by considering *where* average differences in punishment between students of color and their white peers emerge. When we consider incidents that implicate at least two boys for fighting we find that students of color receive harsher punishments, on average—across 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 multiracial 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. Our analysis suggests that biases are still important, and that they can produce large differences in the disciplinary experiences of students. However, at the same time, we find encouragement insofar as the data-generating process supports that biases are correctable where race and equality norms may be more salient.

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	

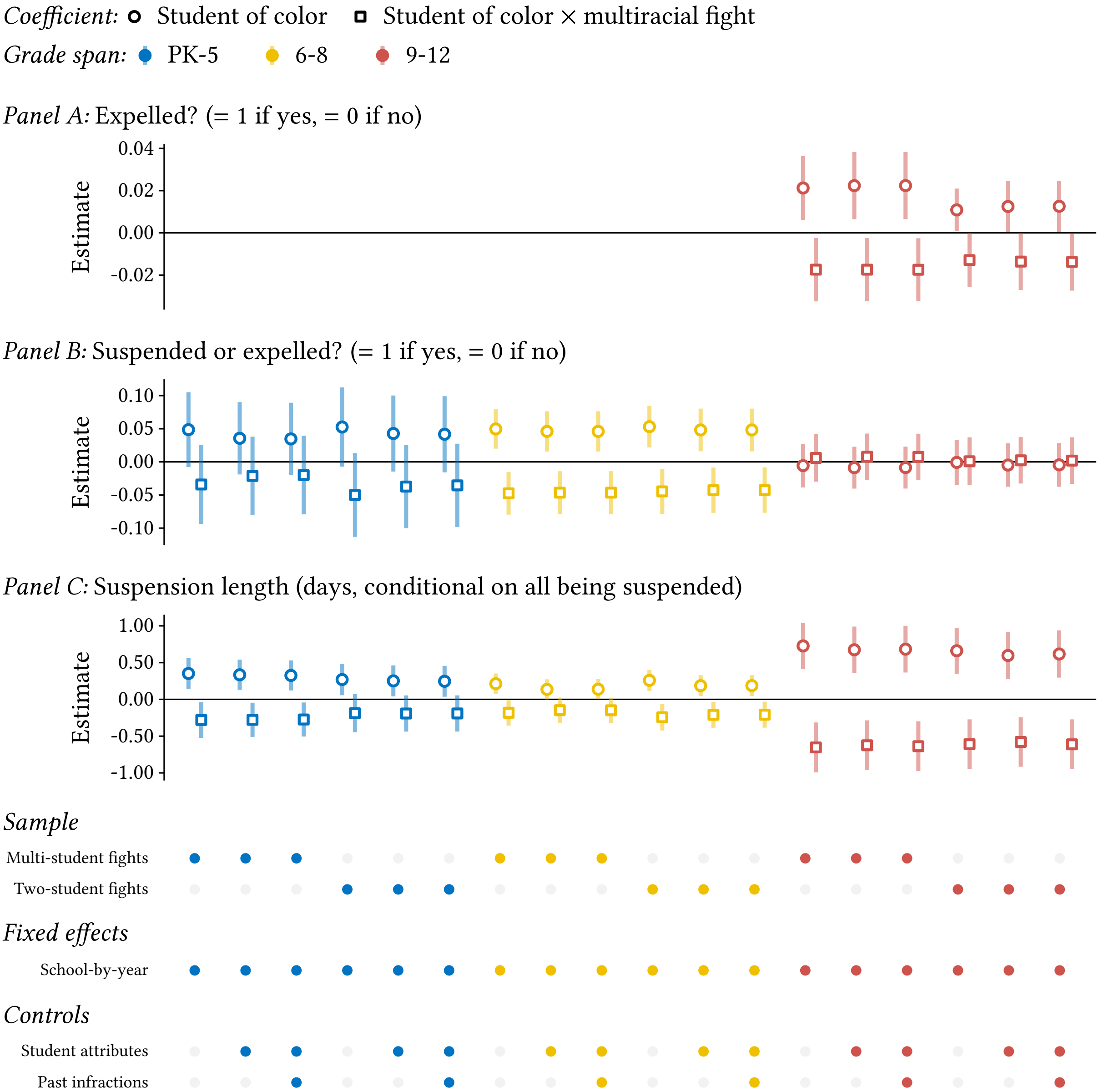
Notes: Sample means (μ) and standard deviations (σ) of punishment outcomes considered in Section 3 and Section 4. The alternative to an expulsion or a suspension is either "no intervention" or "other intervention." Suspension lengths are conditional on all students being suspended within each incident. 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.

References

Anderson, K. P. and 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. and 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., and 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., and 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., and Deming, D. J. (2019). The school to prison pipeline: Long-run impacts of school suspensions on adult crime. Working Paper 26257, National Bureau of Economic Research.

Barrett, N., McEachin, A., Mills, J. N., and Valant, J. (2019). Disparities and discrimination in student discipline by race and family income. *Journal of Human Resources*, pages 0118–9267R2.

Beck, A. N. and 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. and Payne, A. A. (2017). High school choices and the gender gap in STEM. *Economic Inquiry*.

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

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., and 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. and Martin, D. C. (2019). Discipline reform, school culture, and student achievement.

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.

Glenn, B., Barrett, N., and Lightfoot, E. (2020). The effects and implementation of restorative practices for discipline in New Orleans schools. Technical report, Education Research Alliance for New Orleans.

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

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

Lacoe, J. and 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.

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

Michelmore, K. and 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. and Zuo, G. W. (2020). Suspending suspensions: The education production consequences of school suspension policies.

Reykdal, C., Weaver-Randall, K., and Ireland, L. (2018). Comprehensive Education Data and Research System (CEDARS) appendix manual, version 10.2. Technical report, Washington State Office of Superintendent of Public Instruction.

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

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