Understanding a Vicious Cycle: The Relationship Between Student Discipline and Student Academic Outcomes

Kaitlin P. Anderson[1], Gary W. Ritter[2], and Gema Zamarro[3]

While numerous studies have demonstrated a correlation between exclusionary discipline and negative student outcomes, this relationship is likely confounded by other factors related to the underlying misbehavior or risk of disciplinary referral. Using 10 years of student-level demographic, achievement, and disciplinary data from all K–12 public schools in Arkansas, we find that exclusionary consequences are related to worse academic outcomes (e.g., test scores and grade retention) than less exclusionary consequences, controlling for type of behavioral infraction. However, despite controlling for a robust set of covariates, sensitivity checks demonstrate that the estimated relationships between consequences and academic outcomes may still be driven by selection bias into consequence type. Implications for policy and practice are discussed.

Keywords: achievement; at-risk students; educational policy; regression analyses; student behavior/attitude

Recently, many school districts and states have enacted policies to limit suspensions, likely in response to the well-established link between suspensions and undesirable academic outcomes. There is also substantial evidence that out-of-school suspensions (OSSs) and expulsions are disproportionately assigned to certain types of students, particularly students of color (e.g., Anderson & Ritter, 2017; Anyon et al., 2014; Skiba, Chung et al., 2014). However, it is also possible that the relationship between consequences and negative student outcomes is correlational rather than causal. For the sake of crafting good policy, it is important that policymakers gain a better grasp of the true impacts of various types of disciplinary responses.

In this paper, we estimate the relationships between seven types of disciplinary responses to student behavior and two academic outcomes—math test scores and grade retention—while controlling for a rich set of observable characteristics that predict selection into disciplinary referrals and consequences. While estimating a causal relationship is difficult because of the potential for reverse causality or common causes, our detailed data provide a unique opportunity to estimate these relationships. Specifically, our contribution is the ability to control for infraction type, thereby disentangling the impact of the consequence from the underlying factors causing misbehavior.

In the sections that follow, we summarize prior evidence on the relationship between student discipline and student outcomes, describe our data and analytic approach, and discuss our results and their implications for designing policy solutions.

Literature Review

We know little about the causal effect of disciplinary consequences on student outcomes (Steinberg & Lacoe, 2016), yet a large body of evidence has documented correlations between exclusionary discipline (e.g., suspensions and expulsions) and negative student outcomes including lower academic achievement (Arcia, 2006; Beck & Muschkin, 2012; Cobb-Clark, Kassenboehmer, Le, McVicar, & Zhang, 2015; Kinsler, 2013; Noltemeyer, Ward, & Mcloughlin, 2015; Raffaele-Mendez, 2003), school drop-out and grade retention (Balfanz, Byrnes, & Fox, 2014; Carpenter & Ramirez, 2007; Fabelo et al., 2011; Suh & Suh, 2007; Swanson, Erickson, & Ritter, 2017), and involvement in the criminal or juvenile justice systems (Fabelo et al., 2011; Nicholson-Crotty, Birchmeier, & Valentine, 2009; Wolf & Kupchik, 2017).

There are several mechanisms through which academic declines could occur, but the complex relationship between

[1] Michigan State University, East Lansing, MI
[2] Saint Louis University, St. Louis, MO
[3] University of Arkansas, Fayetteville, AR

Educational Researcher, Vol. 48 No. 5, pp. 251–262
DOI: 10.3102/0013189X19848720
Article reuse guidelines: sagepub.com/journals-permissions

behavior, consequences, and academic outcomes leads to uncertainty about the causal relationships. Lower academic achievement could be a result of lost instruction when suspended (Scott & Barrett, 2004). However, students struggling academically exhibit undesirable behaviors in later time periods (e.g., Arcia, 2006; Beck & Muschkin, 2012; Choi, 2007), raising questions about the causal direction. Suspended or expelled students might develop feelings of isolation, stigmatization, or disengagement from school following a suspension, which could translate into poor academic outcomes; however, students may have felt these feelings prior to their exclusion as well (Morrison et al., 2001).

Moreover, it is difficult to isolate the impact of the consequence itself. Most studies have estimated the difference in outcomes for excluded students, relative to nonexcluded students, without direct comparisons to students who behaved similarly but were nonexcluded. For example, a comparison of the relative effects of OSS and in-school suspension (ISS) would be informative for educators and policymakers, as ISS is a commonly used alternative to OSS. However, a literature review noted a lack of research on the effects of ISS (Noltemeyer et al., 2015).

Due to these challenges, the existing literature generally fails to distinguish the impact of punishment from the underlying factors leading to it. In this section, we summarize the literature on the relationship between student discipline, academic achievement, grade retention, and drop-out and discuss the evidence on the impacts of suspension-reducing policies.

Relationship Between Student Discipline and Achievement

A meta-analysis of 24 studies from 1986 to 2012 determined there was a significant, negative relationship between suspensions and academic achievement (Noltemeyer et al., 2015). OSS was more strongly associated with achievement ($r = -0.25$) than ISS was ($r = -0.10$), but the authors do not emphasize this difference because few studies assessed ISS and OSS separately. About 42% of the included studies assessed the correlation between school-level suspension rates and achievement, rather than using student-level data, and most did not control for presuspension academic or behavioral factors (Noltemeyer et al., 2015).

Several student-level studies have found a negative relationship between OSS and academic achievement conditional on demographic and contextual characteristics (Arcia, 2006; Cobb-Clark et al., 2015; Kinsler, 2013; Raffaele-Mendez, 2003). Yet these studies did not control for baseline achievement, leaving an important variable omitted. Indeed, Cobb-Clark et al. (2015) conducted a sensitivity test proposed by Altonji, Elder, and Taber (2005) to assess whether selection bias might be driving their results. They concluded that the estimated relationship between suspensions and academic outcomes is unlikely to be causal and is likely a function of other differences, not controlled for, between suspended and nonsuspended students.

Relationship Between Disciplinary Consequences, Grade Retention, and Drop-Out

Disciplinary issues, including those indicated by the observance of school suspensions, are commonly found to predict both grade retention and drop-out (Balfanz et al., 2014; Carpenter & Ramirez, 2007; Chu & Ready, 2018; Fabelo et al., 2011; Suh & Suh, 2007; Swanson et al., 2017). Next, we describe a few of the recent studies on this topic, focusing on student-level studies in particular.

Suh and Suh (2007), using the National Longitudinal Survey of Youth (NLSY97), found that suspended students were 77.5% more likely to drop-out than nonsuspended students, controlling for factors such as GPA, absenteeism, fighting, receiving threats in school, family structure, socioeconomic status, and school contextual factors. Notably, survey data often provide many more covariates that are typically available in administrative data.

Fabelo et al. (2011) estimated that Texas students suspended or expelled for discretionary violations—for which exclusion is not mandated—relative to their nonexcluded but otherwise similar peers (in terms of demographics and prior achievement, attendance, and past disciplinary issues) in similar schools, were about twice as likely to be retained in grade.

Using 7 years of student- and infraction-level data from Arkansas, Swanson et al. (2017) found that students who receive exclusionary discipline in eighth grade were 2.5 percentage points (PPTs) more likely to be retained in ninth, conditional on demographics, baseline achievement, school fixed effects, and notably, the types of infractions reported.

Chu and Ready (2018) used a within-student approach, comparing student-semesters with and without suspensions and found that students were more likely to drop out the semester following a suspension. Bias would remain in these estimates, however, if an external shock contributed to both suspension and drop-out. In a separate matching analysis, they found that students suspended in the first three semesters of high school were less likely to graduate than their peers with similar demographics, eighth grade test scores, and pre-high school history of absences, tardies, and suspensions, who were not suspended in their first three semesters of high school. They do not control for the types of behaviors that led to suspensions.

An important result across these studies is that the magnitude of the relationship between exclusionary discipline and academic outcomes is often diminished when controlling for student, school, and family characteristics (e.g., Fabelo et al., 2011; Swanson et al., 2017). While a few account for student behavior in some way (Chu & Ready, 2018; Fabelo et al., 2011; Suh & Suh, 2007; Swanson et al., 2017), only one compared exclusionary and nonexclusionary consequences for similar infractions (Swanson et al., 2017), which directly addresses the fundamental policy question of the impact of using exclusionary discipline for a given offense.

Limited Evidence Isolates the Impact of Consequences Separately From Behavior

The studies just described—plus others that focus on outcomes such as criminal involvement (e.g., Wolf & Kupchik, 2017)—demonstrate a statistically significant relationship between discipline and student outcomes after controlling for student demographics and baseline achievement. Thus, in a literature review, Skiba, Arredondo, and Williams (2014) argued that "above and beyond individual, family, and community risk factors, exclusionary school discipline makes a significant contribution in and of itself to a range of negative developmental outcomes" (p. 556). However, with few exceptions (e.g. Swanson et al., 2017), a key variable remains omitted: the misbehavior leading to these disciplinary consequences. Therefore, a key contribution of our study is the ability to compare outcomes for students who receive different disciplinary responses for the same type of infractions.

Suspension-Reducing Policies and Student Outcomes

One strategy to better understand the effects of suspension is to assess what happens when suspensions are banned or limited. A small but growing literature on suspension-reducing policies indicates a mix of benefits and unintended consequences with respect to student outcomes.

For example, in some cases, reducing suspensions resulted in improved attendance and achievement overall (Hinze-Pifer & Sartain, 2018) or for suspended students (Steinberg & Lacoe, 2018). In others, attendance—but not academic achievement—improved following reductions in the length of suspensions (Sartain et al., 2015). Of course, this evidence base is new and growing and there is no consensus. Indeed, some studies have found that reductions in suspensions may also have unintended consequences such as increased truancy and declines in achievement (Lacoe & Steinberg, 2018) as well as deteriorating school climate (Sartain et al., 2015; Lacoe & Steinberg, 2018).

Given the limited and mixed nature of these findings, it is too early to draw conclusions about the overall impacts of these policies, but the current study—which estimates the relationship between disciplinary consequences and academic outcomes while controlling for a uniquely rich set of observable characteristics—sheds some light on what we should expect from suspension-reducing policies.

Data

This study uses 10 years of de-identified demographic, achievement, and disciplinary data from all K–12 public schools in Arkansas provided by the Arkansas Department of Education (ADE) for 2007–08 through 2016–17. Demographic data include race, gender, grade, special education status, limited English proficiency (LEP), and free- and reduced-price lunch (FRL) eligibility. Academic achievement data include scores on state mathematics and reading (ELA) tests in Grades 3 through 8 from 2008–09 to 2016–17. All test scores are standardized by test type, grade level, and academic year to have a mean of zero and standard deviation of one (i.e. *z* scores).[1] The available data do not include indicators of grade retention, but we are able to infer these outcomes based on student grade-level assignments in consecutive years.[2]

Discipline data are provided at the incident level and include indicators for infraction type, consequence type and, in some cases, the length of the consequence in days. There were 1,940,772 infractions during our 10-year panel. Three subjective categories, disorderly conduct (28.7%), insubordination (23.8%), and other nonspecified infractions (28.1%),[3] represent over 80% of infractions. Other types include fighting (6.8%), truancy (6.3%), bullying (2.4%), tobacco (1.2%), student assault (1.0%), drugs (0.6%), vandalism (0.5%), knives (0.2%), staff assault (0.2%), alcohol (0.2%), and gang-related activity (0.1%). Gun, explosives, and club infractions are extremely rare. There are seven consequence types recorded as ISS (37.3%), other nonspecified consequences (27.0%),[4] OSS (21.8%), corporal punishment (12.6%), no action (0.8%), referrals to Alternative Learning Environments (ALE) (0.3%), and expulsions (0.1%). We aggregate disciplinary data to the student-by-academic year level, creating variables that indicate the number of infractions and consequences, by type.

There are, on average, 0.4 infractions per student per year, or 2.7 infractions per student among the student-years with at least one referral. Middle and high school students have more infractions than younger students; 3rd through 5th graders have on average 0.26 infractions per year, 6th through 8th graders average 0.59, and 9th through 12th graders average 0.55.

Methods

We use a variety of specifications to estimate the relationships between each consequence type and two academic outcomes (math test scores and grade retention).[5] We estimate models using disciplinary consequences in the current year (CY), the prior year (PY), or both, to test for persistent relationships over time. We estimate nested models that start with a narrow set of controls and sequentially add more controls, in order to observe the change in the estimated relationships. Test score models are estimated for students in Grades 3–8, in which annual testing occurs. Grade retention models are estimated using students in Grades 9–11, because the risk of discipline-induced drop-out, perhaps due in part to grade retention, is more common in high school.

Our fully specified linear regression model incorporating only CY discipline measures is:

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 y_{it-1} + \boldsymbol{Cons_{it}}\alpha + \boldsymbol{Infrac_{it}}\rho + \tau_{\boldsymbol{t}} \\ + \sigma_{\boldsymbol{s}} + \boldsymbol{X_{it}}\gamma + \varepsilon_{it}. \end{aligned} \quad (1)$$

Our fullest models that incorporate PY discipline measures in lieu of or in conjunction with the CY discipline measures are represented by the following two equations:

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 y_{it-2} + \boldsymbol{Cons_{it-1}}\delta + \boldsymbol{Infrac_{it-1}}\vartheta + \tau_{\boldsymbol{t}} \\ + \sigma_{\boldsymbol{s}} + \boldsymbol{X_{it}}\gamma + \varepsilon_{it}, \end{aligned} \quad (2)$$

$$\begin{aligned} y_{it} = \beta_0 + \beta_1 y_{it-2} + \boldsymbol{Cons_{it}}\alpha + \boldsymbol{Cons_{it-1}}\delta + \boldsymbol{Infrac_{it}}\rho \\ + \boldsymbol{Infrac_{it-1}}\vartheta + \tau_{\boldsymbol{t}} + \sigma_{\boldsymbol{s}} + \boldsymbol{X_{it}}\gamma + \varepsilon_{it}. \end{aligned} \quad (3)$$

In the test score models, we control for lagged measures of the outcome, y_{it-1}, or in equations that include PY discipline measures, we control for twice-lagged versions, y_{it-2}, as a prediscipline baseline measure.[6] For the grade retention outcomes, we estimate discrete choice probit models predicting the probability of grade retention for student *i* in year *t* as a function of the same variables, except y_{it-1} and y_{it-2} are replaced by eighth grade (baseline) test scores, because it is extremely rare for a student to be retained 2 years in a row. For 9th to 11th grade students, of the

2.85% of student-year observations with grade retention, only 7.88% of these experienced grade retention in the PY. The grade retention results are very similar with and without the eighth grade test score controls.[7]

The variables of interest are vectors of consequence counts, $\boldsymbol{Cons}_{it}$ and $\boldsymbol{Cons}_{it-1}$. These vectors each include six variables indicating the count of incidences of six types of consequences (expulsions, referrals to an ALE, ISS, corporal punishment, "other" nonspecified consequences, and no action) in the same and PY, respectively, with OSS consequences as the reference category. For these count variables, each incidence is counted once, regardless of the number of days associated with it.

Two vectors, $\boldsymbol{Infrac}_{it}$ and $\boldsymbol{Infrac}_{it-1}$, include incident counts for each of 17 infraction types, which allow us to control for the type of misbehavior leading to disciplinary consequences. A remaining limitation is that infraction types include a range of behaviors, and the resulting consequence type is likely related to underlying—but unobservable—characteristics of the misbehavior. For example, within the disorderly conduct category, more severe behaviors might result in more exclusionary consequences and may also be associated with worse academic outcomes regardless of consequence type. Inconsistent reporting practices within or across schools could also bias the results, although we do use school fixed effects to address such differences—and all other time-invariant characteristics—across schools. In addition, given that our outcome measures are annual measures, we are not estimating the outcomes associated with each particular incident. Rather, we estimate the correlation between the cumulative set of consequences, controlling for the cumulative set of infractions in the same year. Thus, we are not accounting for heterogeneity of the estimated relationships based on the other reported misbehavior that year, which may be something that school administrators consider when determining consequences.[8] Despite these limitations, our approach represents an improvement upon prior work that generally does not control for the behavior leading to suspensions.

The coefficients should be interpreted differently in the models with and without infraction controls. In the models that control for infraction counts, the infraction controls and consequence controls are perfectly collinear, because the total number of infractions equals the total number of consequences. In these models, the OSS consequence is dropped and used as the reference category, and as a result, we estimate the "impact" of six consequence types, relative to receiving an OSS, for students referred for the same infraction types. We use OSS as the reference category because it is most commonly the focus of research and political discussions surrounding discipline reform. In contrast, in the models that do not include these infraction controls, there is a seventh variable indicating the count of OSS consequences. In these models, we estimate the "impact" of seven consequence types (including OSS), without accounting for reported behavior.

We account for school-level time-invariant characteristics with school fixed effects, $\boldsymbol{\sigma}_s$, and for state-wide differences over time using academic year indicators, $\boldsymbol{\tau}_t$. We control for student characteristics, X_{it}, including binary indicators of gender, FRL status, special education status, LEP status, race/ethnicity (Black, Hispanic, Asian, and other, with White as the reference group), and grade levels. There may be remaining endogeneity concerns due to omitted variables, so we do not interpret these estimates as causal, but rather as the relationships between disciplinary consequences and student outcomes, controlling for a uniquely rich set of covariates including type of behavior reported. These relationships are directly relevant to discussions surrounding discipline policy reform.

One concern with administrative discipline datasets is the underreporting of infractions that do not result in suspensions or expulsions. Even though we do not know how many instances of misbehavior are not reported, we do at least observe a variety of consequences beyond simply suspensions or expulsions. Less than a quarter of reported infractions result in the most exclusionary consequences (expulsions, ALE referrals, and OSS), 37% result in ISS, and about 40% result in other types of consequences (see Table 1).[9] Some infractions such as truancy seldom result in OSS, but for others, like drugs or alcohol, students receive OSS in almost 90% of cases. Within each infraction type, there is variation in disciplinary response, and several infraction types resulted in a relatively even mix of OSS and non-OSS consequences.[10] The use of school-fixed effects helps control for differences in reporting patterns across schools.

Descriptive statistics for the full state and four key analytic samples are in Table 2. The analytic samples are generally similar to the entire state, except that students are less likely to be FRL-eligible or LEP in the grade retention sample, which only includes Grades 9–11.

In Table 3, we provide descriptive statistics for five groups of student-academic year observations: all observations, those with any infractions, those with any exclusionary discipline, those with at least one ISS, and those with at least one OSS. Disciplined students, particularly excluded students, are more likely to be older, non-White, FRL-eligible, receiving special education services, lower performing, and retained in grade than the general student population. The first column shows the relative rarity of exclusionary discipline in general. The average student has 0.408 infractions and 0.089 OSS incidents per year.

Results

Relationship Between Discipline and Student Achievement

Table 4 shows the relationship between disciplinary consequences and math test scores. Recall that the coefficients in the models with infraction controls (columns 2, 4, 6, 8, and 9) should be interpreted as the relationship between consequences and math test scores, relative to the relationship between an OSS consequence and math test scores, for the same reported infraction(s). This differs from columns 1, 3, 5, and 7, which estimate the relationship between consequences and outcomes without controlling for infractions. In columns 1–4, CY measures of disciplinary outcomes are used; in columns 5–8, PY measures of disciplinary outcomes are used; and in column 9, both CY and PY measures are included.

The results in columns 1, 3, 5, and 7 indicate a consistently negative or null relationship between counts of consequences and test scores. The largest relationships are between expulsions and test scores, and the magnitude tends to decline as the severity or degree of exclusion declines. Columns 2, 4, 6, 8, and 9 also demonstrate that more exclusionary consequences are associated with lower test scores. For example, in these columns, which use

Table 1
Percent of Incidents Resulting in Various Consequences, by Infraction Type (K–12)

		Percent of Infractions Resulting in Each Consequence Type						
	Total Number	Expulsion	ALE	OSS	ISS	Corp. Pun.	Other	No Action
Disorderly conduct	556,790	0.0%	0.3%	19.2%	33.0%	13.6%	33.0%	0.8%
Other	544,709	0.1%	0.2%	16.9%	37.2%	15.9%	28.9%	0.9%
Insubordination	461,445	0.0%	0.2%	17.3%	40.8%	12.9%	27.8%	0.9%
Fighting	131,699	0.2%	0.3%	60.1%	24.4%	7.2%	7.6%	0.2%
Truancy	121,787	0.0%	0.1%	11.1%	62.6%	2.5%	23.0%	0.6%
Bullying	45,961	0.1%	0.2%	25.9%	40.7%	11.9%	20.4%	0.7%
Tobacco	22,645	0.1%	0.4%	33.5%	48.9%	7.8%	9.3%	0.1%
Student assault	20,135	0.5%	0.9%	47.5%	27.5%	13.4%	10.0%	0.3%
Drugs	11,327	3.3%	1.3%	85.1%	8.5%	0.1%	1.7%	0.1%
Vandalism	9,799	0.2%	0.3%	30.1%	36.7%	9.4%	22.7%	0.6%
Knife	4,347	2.2%	1.7%	72.1%	17.1%	2.0%	4.7%	0.1%
Staff assault	3,669	1.1%	5.5%	68.1%	9.2%	6.5%	9.3%	0.3%
Alcohol	3,174	1.2%	0.9%	89.3%	7.3%	0.1%	1.1%	0.0%
Gangs	2,008	1.0%	0.7%	61.8%	22.5%	0.5%	12.6%	0.9%
Explosives	471	0.6%	0.8%	47.8%	27.6%	5.3%	17.6%	0.2%
Guns	430	7.7%	4.0%	46.3%	17.2%	20.9%	4.0%	0.0%
Club	376	1.1%	10.4%	72.1%	8.2%	0.5%	7.4%	0.3%
Total	1,940,772	0.1%	0.3%	21.8%	37.3%	12.6%	27.0%	0.8%

Note. ALE = referral to Alternative Learning Environment; Corp. Pun. = corporal punishment; ISS = in-school suspension; OSS = out-of-school suspension. "Other" categories for both infractions and consequences refer to incidents that were not included in one of the state reporting categories and are not researcher-created categories.

Table 2
Descriptive Statistics for State and Analytic Samples

	Entire State	CY Math	Fully Specified Math	CY Grade Retention	Fully Specified Grade Retention
N observations	4,759,356	1,353,471	913,958	536,227	532,417
N students	950,745	445,294	350,529	220,445	219,109
Male	51.3%	51.1%	51.2%	50.6%	50.6%
FRL	60.3%	62.1%	61.8%	54.6%	54.6%
Special education	11.4%	11.6%	11.5%	10.3%	10.3%
Limited English proficient	7.3%	7.6%	7.5%	5.2%	5.2%
White	64.3%	64.1%	64.0%	65.3%	65.3%
Black	21.2%	20.8%	20.9%	21.6%	21.6%
Hispanic	10.4%	10.9%	10.9%	9.4%	9.4%
Other race	4.1%	4.3%	4.3%	3.8%	3.8%
Average grade level	5.9	6.0	6.5	9.8	9.8
Math *z* score	0.000	0.015	0.016	0.049	0.050
ELA *z* score	0.000	0.008	0.007	0.039	0.040

Note. Descriptive statistics for student-by-year observations. Current year (CY) samples refer to columns 3 and 4 in Tables 4 and 5. The "fully specified" samples refer to columns 7–9 in Table 4 and columns 7, 8, and 10 in Table 5. The math and ELA test scores reported in the grade retention samples refer to eighth grade (baseline) test scores. FRL = free- and reduced-price lunch; ELA = English language arts.

OSS as the reference category, the coefficients on expulsion and ALE are generally negative, while the coefficients on less exclusionary consequences are generally positive. After controlling for infraction types and lagged math test scores (column 4) and relative to an OSS consequence, each expulsion is associated with –0.103 *SD* lower math test scores, each ISS incident is associated with 0.013 *SD* higher scores, and each "other" (generally nonexclusionary) consequence is associated with 0.026 *SD* higher scores. Thus, there is a clear relationship between degree of exclusion and achievement outcomes.

Notably, without controlling for student behavior (columns 1, 3, and 5), even "no action/warning" has a negative relationship to test scores, demonstrating the importance of infraction controls for drawing conclusions about the impact of

Table 3
Descriptive Statistics for Student-Year Observations, by Level of Discipline Exposure

	All Observations	Any Infractions	Any Exclusions	At Least 1 ISS	At Least 1 OSS
N observations	4,759,356	716,321	254,101	358,033	251,515
N students	950,745	329,723	146,898	197,838	145,738
Male	51.3%	67.8%	69.1%	67.4%	69.1%
FRL	60.3%	74.1%	79.9%	74.7%	80.0%
Special education	11.4%	16.1%	18.6%	16.2%	18.6%
Limited English proficient	7.3%	4.9%	4.0%	5.7%	4.0%
White	64.3%	51.5%	38.0%	49.5%	37.9%
Black	21.2%	37.5%	53.3%	38.2%	53.4%
Hispanic	10.4%	8.0%	6.3%	9.1%	6.3%
Other race	4.1%	3.0%	2.4%	3.2%	2.4%
Average grade level	5.86	7.02	7.35	7.66	7.35
Discipline per student per year					
N infractions	0.408	2.709	3.734	3.594	3.746
N expulsions	0.000	0.003	0.007	0.002	0.003
N ALE referrals	0.001	0.009	0.024	0.008	0.017
N OSS	0.089	0.591	1.665	0.528	1.682
N ISS	0.152	1.012	1.117	2.024	1.120
N corporal punishments	0.052	0.343	0.247	0.231	0.247
N no actions	0.006	0.042	0.035	0.037	0.035
N "other" consequences	0.107	0.711	0.638	0.764	0.641
Count of 3rd–8th grade obs.	2,228,927	347,041	127,795	176,404	126,626
Math *z* score (Grade 3–8)	0.000	–0.464	–0.671	–0.514	–0.671
ELA *z* score (Grade 3–8)	0.000	–0.513	–0.720	–0.571	–0.720
Count of 9th–11th grade obs.	1,081,423	225,625	86,290	129,457	85,325
Grade retention (Grade 9–11)	2.9%	6.7%	11.6%	6.9%	11.6%

Note. Descriptive statistics are based on student-year observations (obs.) over the full panel and include all grades unless otherwise specified. For the outcome variables, we limit the observations to those for Grades 3–8 (test score outcomes) or 9–11 (grade retention outcome). The counts for these 3–8 and 9–11 grade outcomes are the count of all student observations in those grades, even if they did not have the outcome variable. ISS = in-school suspension; OSS = out-of-school suspension; FRL = free- and reduced-price lunch; ALE = referral to Alternative Learning Environment; ELA = English language arts.

consequences. Similarly, the estimated adverse relationships are greatly diminished after including baseline test scores, indicating the importance of baseline characteristics. Thus, a key takeaway from Table 4 is the importance of controlling for reported infraction type and baseline achievement when assessing the potential effects of consequences on academic achievement.

Further, there do appear to be some persistent relationships between test scores and disciplinary consequences in the PY. For example, relative to OSS in the PY, ALE in the PY is associated with lower test scores, and some nonexclusionary consequences in the PY (ISS and "other") are associated with higher test scores. Notably, when adding the PY measures (column 9), the point estimates on the CY measures change very little (relative to in column 4). We may be underestimating the importance of PY consequences if they produce future bad behavior, as this relationship would be captured in our controls for CY discipline. On the other hand, it is also possible that remaining unobservables are driving part of these relationships, which appear persistent over time.

Relationship Between Discipline and Grade Retention

The results of our grade retention models are in Table 5. In columns 1–4, CY measures of disciplinary outcomes are used; in columns 5–8, PY measures are used; and in columns 9–10, both CY and PY measures are included. Columns 3, 4, 7, 8, and 10 control for eighth grade test scores. The models that do not control for reported infraction types (columns 1, 3, 5, and 7) generally indicate that exclusionary consequences such as expulsions, referrals to ALE, OSS, and ISS are all associated with higher risk of grade retention. "Other" consequences are sometimes associated with lower risk of grade retention. Columns 2, 4, 6, and 8–10 control for the types of infractions reported. These results indicate that more exclusionary consequences like expulsion and ALE—particularly in the CY—are generally associated with a higher likelihood of grade retention, relative to OSS. On the other hand, ISS, corporal punishment, no action/warning, and "other" consequences are associated with lower risk, relative to OSS, indicating, as in Table 4, that less exclusionary consequences have a weaker association with negative academic outcomes. To interpret the size of these coefficients, it is important to note that grade retention is quite rare. Only 2.85% of student-year observations in Grades 9–11 indicated grade retention, and so a 0.7 PPT increase in the likelihood of grade retention for each OSS incident (as in column 3) represents a 25% increase, a large effect. Columns 9–10 suggest that PY consequences are also predictive of grade retention, independent of CY consequences, with all the

Table 4
Relationship Between Disciplinary Consequences and Math Test Scores in Grades 3–8

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lagged math *z* score			0.704***	0.704***					
			(0.004)	(0.004)					
Twice lagged math *z* score							0.672***	0.672***	0.668***
							(0.006)	(0.006)	(0.006)
CY expulsion	−0.269***	−0.132***	−0.160***	−0.103***					−0.108**
	(0.038)	(0.040)	(0.033)	(0.035)					(0.051)
CY ALE	−0.155***	−0.063**	−0.069***	−0.036*					−0.041*
	(0.023)	(0.024)	(0.020)	(0.021)					(0.025)
CY OSS	−0.107***		−0.040***						
	(0.004)		(0.002)						
CY ISS	−0.060***	0.034***	−0.023***	0.013***					0.012***
	(0.005)	(0.006)	(0.002)	(0.003)					(0.003)
CY corp. pun.	−0.058***	0.041***	−0.023***	0.012***					0.016***
	(0.007)	(0.007)	(0.004)	(0.004)					(0.004)
CY no action/warning	−0.043***	0.061***	−0.023***	0.015*					0.013
	(0.013)	(0.011)	(0.008)	(0.008)					(0.010)
CY other consequence	−0.026***	0.072***	−0.010***	0.026***					0.023***
	(0.005)	(0.006)	(0.002)	(0.003)					(0.004)
PY expulsion					−0.130***	−0.016	−0.025	0.024	0.014
					(0.044)	(0.044)	(0.045)	(0.047)	(0.046)
PY ALE					−0.134***	−0.045**	−0.073***	−0.043**	−0.034*
					(0.022)	(0.022)	(0.019)	(0.018)	(0.019)
PY OSS					−0.101***		−0.039***		
					(0.004)		(0.003)		
PY ISS					−0.057***	0.039***	−0.022***	0.013***	0.007**
					(0.004)	(0.005)	(0.003)	(0.004)	(0.003)
PY corp. pun.					−0.061***	0.035***	−0.032***	0.003	−0.007**
					(0.005)	(0.005)	(0.003)	(0.004)	(0.003)
PY no action/warning					−0.022*	0.071***	−0.012	0.022*	0.011
					(0.013)	(0.013)	(0.011)	(0.011)	(0.011)
PY other consequence					−0.025***	0.069***	−0.011***	0.024***	0.010***
					(0.004)	(0.005)	(0.002)	(0.003)	(0.003)
CY inf. counts	N	Y	N	Y	N	N	N	N	Y
PY inf. counts	N	N	N	N	N	Y	N	Y	Y
Constant	0.424***	0.423***	0.588***	0.587***	0.436***	0.434***	−0.173	−0.176	−0.190
	(0.010)	(0.010)	(0.018)	(0.018)	(0.010)	(0.010)	(0.269)	(0.269)	(0.268)
Observations	1,936,923	1,936,923	1,353,471	1,353,471	1,841,902	1,841,902	913,958	913,958	913,958
Adjusted *R*-squared	0.268	0.268	0.639	0.639	0.271	0.271	0.604	0.604	0.606

Note. Robust standard errors, clustered at the school level, are in parentheses. All models include school fixed effects, academic year fixed effects, grade level fixed effects, and student demographic controls including gender, FRL status, special education status, LEP, and a vector of race/ethnicity indicators (White, Black, Hispanic, Asian, and Other). CY and PY infraction (inf.) counts are vectors of variables representing the number of infractions of each type in the current year (CY) or prior year (PY).
$^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$. ALE = referral to Alternative Learning Environment; OSS = out-of-school suspension; ISS = in-school suspension; corp. pun. = corporal punishment; FRL = free- and reduced-price lunch; LEP = limited English proficiency.

less exclusionary consequences having a statistically significant difference, relative to OSS. As in Table 4, when adding the PY measures (column 10), the point estimates on the CY measures change very little (relative to in column 4).

Relative to Table 4, the inclusion of baseline test score measures does not change the point estimates as much (e.g., comparing columns 1 and 3 and comparing columns 2 and 4). Including eighth grade test scores greatly diminishes our sample size, because a significant proportion (18.73%) of students enter our dataset after eighth grade. Therefore, given the similarity between the results with and without eighth grade scores, for additional tests, we focus on the broader sample that does not require eighth grade scores.

Assessing Remaining Selection Bias

To assess whether selection bias may remain, we conduct sensitivity tests proposed by Altonji et al. (2005) and Oster (2017).

Table 5
Relationship Between Disciplinary Consequences and Grade Retention, Grades 9–11

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
8th grade math score			−0.0124***	−0.0123***			−0.0124***	−0.0123***		−0.0118***
			(0.0004)	(0.0004)			(0.0004)	(0.0004)		(0.0004)
8th grade ELA score			−0.0086***	−0.0087***			−0.0086***	−0.0086***		−0.0082***
			(0.0003)	(0.0003)			(0.0003)	(0.0003)		(0.0003)
CY expulsion	0.0287***	0.0140***	0.0204***	0.0107***					0.0108**	0.0102**
	(0.0048)	(0.0046)	(0.0041)	(0.0041)					(0.0046)	(0.0043)
CY ALE	0.0093*	0.0008	0.0106***	0.0050**					0.0001	0.0035
	(0.0055)	(0.0040)	(0.0032)	(0.0025)					(0.0034)	(0.0030)
CY OSS	0.0104***		0.0068***							
	(0.0006)		(0.0004)							
CY ISS	0.0051***	−0.0054***	0.0033***	−0.0036***					−0.0035***	−0.0026***
	(0.0005)	(0.0007)	(0.0004)	(0.0005)					(0.0006)	(0.0004)
CY corp. pun.	0.0007	−0.0088***	−0.0003	−0.0065***					−0.0065***	−0.0055***
	(0.0009)	(0.0011)	(0.0005)	(0.0006)					(0.0008)	(0.0006)
CY no action/warning	0.0004	−0.0098***	−0.0009	−0.0089***					−0.0073***	−0.0081***
	(0.0017)	(0.0025)	(0.0024)	(0.0030)					(0.0026)	(0.0030)
CY other cons.	−0.0019***	−0.0114***	−0.0013***	−0.0075***					−0.0084***	−0.0061***
	(0.0005)	(0.0007)	(0.0003)	(0.0005)					(0.0006)	(0.0005)
PY expulsion					0.0172***	0.0026	0.0109***	0.0003	0.0036	0.0017
					(0.0037)	(0.0037)	(0.0040)	(0.0039)	(0.0043)	(0.0042)
PY ALE					0.0075**	−0.0006	0.0027	−0.0028	0.0002	−0.0026
					(0.0037)	(0.0028)	(0.0025)	(0.0027)	(0.0029)	(0.0027)
PY OSS					0.0092***		0.0065***			
					(0.0005)		(0.0003)			
PY ISS					0.0043***	−0.0045***	0.0029***	−0.0034***	−0.0033***	−0.0025***
					(0.0004)	(0.0006)	(0.0003)	(0.0004)	(0.0005)	(0.0004)
PY corp. pun.					0.0013**	−0.0069***	0.0009**	−0.0049***	−0.0045***	−0.0029***
					(0.0006)	(0.0007)	(0.0004)	(0.0004)	(0.0006)	(0.0005)
PY no action/warning					−0.0007	−0.0098***	0.0000	−0.0069***	−0.0078***	−0.0054***
					(0.0013)	(0.0017)	(0.0016)	(0.0017)	(0.0018)	(0.0017)
PY other cons.					−0.0004	−0.0087***	−0.0002	−0.0062***	−0.0057***	−0.0041***
					(0.0004)	(0.0006)	(0.0004)	(0.0005)	(0.0005)	(0.0005)
CY infraction counts		Y		Y					Y	Y
PY infraction counts						Y		Y	Y	Y
Observations	874,689	874,689	536,227	536,227	749,280	749,280	532,417	532,417	749,280	532,417

Note. Robust standard errors, clustered at the school level, are in parentheses. All models include school, academic year, and grade fixed effects and controls for student gender, free- and reduced-price lunch (FRL) status, special education status, limited English proficiency (LEP) status, and race/ethnicity (White, Black, Hispanic, Asian, and Other). Current year (CY) and prior year (PY) infraction counts are the number of each of 17 infraction types in the current or prior year, respectively. ELA = English language arts; ALE = referral to Alternative Learning Environment; OSS = out-of-school suspension; ISS = in-school suspension; corp. pun. = corporal punishment; cons. = consequences.
$^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Altonji et al. (2005) propose the use of an estimation of the degree of observed selection on observable characteristics as a source of information about the potential selection on unobservables. In particular, their proposed method allows for the estimation of the ratio of selection on unobservables to selection on observables that would be required to attribute the entire estimated effect to selection bias. An assumption required in the case of Altonji et al.'s (2005) test is that if all unobservables were actually observed, the outcome variable could be fully explained (i.e. an *R*-squared of 1), which Oster (2017) argues is not reasonable in the presence of measurement error. Thus, she proposes testing the robustness to the results to alternative assumptions about the maximum *R*-squared possible.

After conducting these tests,[11] we find that we cannot rule out the existence of significant remaining bias, creating uncertainty about whether the estimated relationship is the causal impact of suspension or the result of reverse causality, other confounds, or a mix thereof. For the test score models, the amount of selection on unobservable characteristics would only have to be about 0.26 to 0.56 times as large as the degree of observed selection on observable characteristics to explain all of the estimated effects of expulsion, relative to OSS. To explain all of the estimated effects of ISS on test scores, relative to OSS, selection on unobservables would only have to be about 0.02 to 0.05 times as large as selection on observables. Selection bias in the grade retention models appears even more problematic. Thus, similarly to Cobb-Clark et al. (2015), while this does not prove our estimates are necessarily biased, we cannot rule out the possibility that our observed effects, even after including infraction type controls, are due to remaining selection on unobservables and that true causal effects are actually negligible.

Subgroup Heterogeneity

Despite a lack of support for a causal interpretation, we hypothesize that relationships might differ for certain groups of students if their family background, prior achievement, language proficiency, or disability status make it difficult to make up for lost instructional time. Therefore, we estimated separate results for FRL/non-FRL, White/non-White, LEP/non-LEP, special/regular education, and students whose first test score was above/below average, focusing on the models including CY measures of consequences, with infraction controls.[12]

In general, there are not large differences across subgroups with respect to the relationship between disciplinary consequences and math test scores. In particular, the estimates on corporal punishment and other action were quite similar across subgroups. One notable difference is that ISS is estimated to be potentially more beneficial for relatively disadvantaged groups, when compared to OSS.

Similarly, in our grade retention models by subgroup, the likelihood of grade retention for relatively disadvantaged groups is more sensitive to consequence type. The higher risk associated with expulsion and the lower risk associated with ISS and “other” consequences, relative to receiving OSS, are larger in magnitude for relatively disadvantaged subgroups. If these relationships are causal, this might suggest that the population of students at-risk for grade retention—who are also generally at higher risk of exclusionary discipline—is relatively small but also particularly sensitive to the choice of disciplinary consequence. These findings are consistent whether or not eighth grade math and reading test scores are controlled for.

Robustness of Results Within Restricted Samples of Ever-Disciplined and Ever-Excluded Students

It is possible students who are never disciplined or excluded for disciplinary reasons may affect our estimates if the relationships between covariates and test scores are different for this set of students. Therefore, we re-estimate our models using ever-disciplined students and ever-excluded students (those who received expulsion, OSS, or ALE referral). The results are largely robust to these different samples, except that for the grade retention models, the estimated differences between OSS and other consequences were largest in the full-sample.[13]

Estimate of Nonlinear Relationships

We also tested whether the first exclusionary consequence has a different impact than subsequent consequences by replacing $\boldsymbol{Cons}_{it}$ and $\boldsymbol{Cons}_{it-1}$ with a vector of binary indicators for whether the student had at least one, at least two, at least three, or four or more of that consequence type during the year, with zero incidents as the reference category. We do not draw conclusions about expulsions and ALE for which multiple incidents are extremely rare. For other consequences, the first incident generally has the largest relationship to academic outcomes, suggesting that focusing on preventative strategies and reducing exclusionary discipline for first infractions may be effective.

Discussion and Conclusions

In light of the concerns that exclusionary discipline harms students academically, we set out to generate a better understanding of the magnitude of and the nature of the relationship (e.g., causal or correlational) between disciplinary responses and academic outcomes after controlling for selection into disciplinary consequences. This work makes a key contribution to the field’s knowledge about the potential impact of disciplinary consequences: Controlling for reported infractions is important when estimating the relationship between disciplinary consequences and student outcomes. In addition, controlling for baseline test scores is also important, particularly when estimating the relationship between consequences and test scores.

Not surprisingly, the controls for baseline test scores were more influential in models predicting test scores as outcomes rather than for the models predicting grade retention. Why might the sensitivity of results to the inclusion of baseline test scores as controls vary across academic outcomes? Different results could be due to differences in the analytic samples, grade level in particular (e.g. Grades 3–8 for the test score outcomes and Grades 9–11 for grade retention). Further, serial correlation in test scores is stronger than the correlation between test scores and grade retention, as grade retention is particularly rare. Another possibility is that low achieving students may be more likely to misbehave and perform poorly on standardized tests regardless of the type of punishment (exclusionary or not) but that grade retention is more likely if the resulting punishment is exclusionary, perhaps resulting in lower attendance and failure to complete a course.

Even with the inclusion of important controls such as baseline test scores and behavioral infractions, we still have concerns about remaining selection bias. Specifically, in both our math achievement and grade retention models, the results of sensitivity tests proposed by Altonji et al. (2005) and Oster (2017) suggest there may be bias remaining in our estimated effects with full controls.

Overall, our results indicate that the choice of disciplinary consequence is not driving the entire decline in academic progress, and it is important to consider the relative influence of various approaches in response to student behavior. The recent policy focus on limiting suspensions might miss the mark if there is little guidance about appropriate alternatives. For example, if instead of suspensions, schools refer more students to ALE, this might be a more costly and potentially more harmful approach. Further, simply replacing OSS with ISS may not eliminate the academic decline of disciplined students, because even ISS is associated with negative academic outcomes for students.

Less exclusionary consequences such as those in the "other" consequence category (e.g., Saturday school, detentions, parent/guardian conferences) are associated with better outcomes, relative to both OSS and ISS, although future work is needed to understand more about which specific approaches schools are using and having success with.

Another key finding is that the first disciplinary consequence in a given year is associated with larger academic declines. This implies that policies should target preventative approaches, rather than waiting until students get into major trouble before getting involved. Supporting this idea, many scholars have argued for more proactive discipline focusing on preventing misbehavior by setting clear expectations and teaching students prosocial behaviors (Chin, Dowdy, Jimerson, & Rime, 2012; Sharkey & Fenning, 2012).

Another important takeaway is that, particularly in terms of grade retention, the association between suspensions, expulsions, and negative academic outcomes may be larger for students from historically underserved backgrounds. Given that these students are at elevated risk of being suspended or expelled, this has direct implications for educational attainment gaps.

A few limitations remain. First, we do not estimate the potential impacts on all adverse student outcomes (e.g., we do not link these data to court records to test hypotheses related to the school-to-prison pipeline or assess impacts on student attendance or drop-out), nor do we test the specific mechanisms through which these impacts might occur. In addition, we do not estimate the impacts of suspensions on the nonsuspended students in the school.

Second, we use administrative data that may include attrition or measurement error. With respect to attrition, we observe that students expelled or referred to ALE are slightly more likely to drop out of the Arkansas public school dataset altogether, which may mean that we are slightly underestimating the negative consequences for some students, if the students who were most harmed dropped out or sought other schooling options.[14] Another attrition-related concern is that students who we report as being retained in grade may have effectively dropped out, despite being enrolled for a short period of time. We found, however, that the results were generally similar when dropping students for whom this could possibly be the case.[15]

With respect to measurement error, administrative data only indicate a category of offense, but the underlying behavior of one student's insubordination, for example, may be very different from another's. If anything, we expect that unobservables—such as severity of offense within infraction type—would be correlated with the risk of exclusionary discipline and student academic outcomes in a way that would bias our estimated relationships upward in magnitude. For example, within the category of disorderly conduct, if we assume that students committing worse offenses tend to receive more exclusionary consequences and have worse academic outcomes, we would be overstating the magnitude of the relationship between exclusion and academic outcomes. Thus, we might view our estimates of the relationships between exclusionary consequences and student outcomes as upper bounds on the true relationships. We also use school fixed effects to help account for differences in reporting patterns or policies across schools that are stable over time.

Further, there are likely unobservable characteristics of students that are associated with their risk of discipline and academic outcomes, and earlier versions of this work found that accounting for student heterogeneity greatly attenuated the estimated relationships.[16]

Finally, we rely on school district reports of discipline, and some incidents never make it into the system. This is the case for any analysis using this type of administrative data. A potential area for future research would be ethnographic research to determine the extent to which misbehaviors are simply underreported and how this varies by type of student, teacher, or school.

This study provides a novel look at the impacts of disciplinary responses. Our key contributions are the ability to compare the relationships between various types of consequences and student outcomes, controlling for a unique set of covariates, including reported behavioral infractions, as well as conducting tests for remaining bias, following Altonji et al. (2005) and Oster (2017). To our knowledge, only one study of the relationship between exclusionary discipline and student outcomes was able to control for the particular infractions leading to consequences (Swanson et al., 2017), and only one (Cobb Clark et al., 2015) applied the Altonji et al. (2005) test.

Going forward, as states or districts consider discipline policy reforms, there is a compelling argument for studying the impact of such reforms at the same time. For example, policies aimed at reducing suspensions should consider what the appropriate counterfactual response should be. While there is some experimental evidence supporting the use of School-Wide Positive Behavioral Interventions and Supports (Bradshaw, Mitchell, & Leaf, 2010; Horner et al., 2009) and nonexperimental evidence suggesting the benefits of restorative justice (Fronius, Persson, Guckenberg, Hurley, & Petrosino, 2016), overall, there is little causal evidence on the effectiveness of alternative disciplinary approaches (Steinberg & Lacoe, 2016). Further, interventions such as Positive Behavioral Interventions and Supports (PBIS; Kaufman et al., 2010; Vincent & Tobin, 2011) and restorative justice (Hashim, Strunk, & Dhaliwal, 2018) do not necessarily eliminate racial disproportionalities in suspension, and some suspension-reducing policies also have been implemented inequitably (Anderson, 2018). Thus, evaluations of new programs or policies should address the potential for unintended outcomes as well.

ORCID ID

Kaitlin P. Anderson https://orcid.org/0000-0002-8445-7352

NOTES

[1] From 2008–09 to 2013–14, state tests were administered as part of the Arkansas Comprehensive Testing, Assessment, and Accountability Program (ACTAAP). In 2014–15, Arkansas administered the Partnership for Assessment of Readiness for College and Careers (PARCC) exam, and in 2015–16 and 2016–17, Arkansas administered the ACT Aspire tests. To test the sensitivity of our results to these testing administration changes, we estimated additional models using only the six ACTAAP years, and the results are generally robust. Results are available by request.

[2] Observations from the final study year, 2016–17, are dropped from the grade retention analyses, as without a future year of data it is

impossible to infer retention. Similarly, 12th graders are removed from these analyses as it would be difficult to distinguish between two counterfactuals to grade retention: graduation and drop-out. Dropping 12th graders allows us to identify grade retentions more consistently across grade levels.

[3]"Other" nonspecified infractions were coded as a specific infraction type at the school level but were grouped into an "other" category when reported by the ADE. This is not a researcher-created category.

[4]"Other" nonspecified consequences were coded as a specific consequence type at the school level but were grouped into an "other" category when reported by the ADE. This is not a researcher-created category. Conversations with the ADE Assistant Commissioner for Research and Technology, Eric Saunders, indicated that the majority of these other consequences are detentions, bus suspensions, parent/guardian conferences, Saturday school, or warnings. In fact, in 2016–17, the state started separately reporting additional categories, and in that year at least, 29% were detention, 13% were warnings, 6.2% were Saturday school, 4.5% were bus suspensions, 2.4% were parent/guardian conferences, and the rest were still nonspecified. This reiterates that this category is comprised of predominantly nonexclusionary consequences.

[5]We also estimate effects on reading/English language arts test scores, and the results were generally similar to the math results. Results are available by request.

[6]We also test a variety of specifications including (a) lagged versions of both test scores, (b) lagged and twice lagged versions of the same subject test score, and (c) lagged and twice lagged versions of both test scores. The results are generally robust to these various specifications.

[7]The results are largely robust to the inclusion of this control, but because inclusion of this control greatly diminishes the sample size, we generally focus on models predicting grade retention without the eighth grade test scores.

[8]However, we do test for heterogeneous or nonlinear relationships based on whether the consequence was the first, second, third, or fourth or more for the student in that year and estimate the first infraction to be more highly associated with negative outcomes.

[9]See Note 4 for more detail on these "other" nonspecified consequences.

[10]Table 1 shows the results for all infractions across all grades, but there are some differences based on grade level. For example, relative to the full sample, students in Grades 9–12 were less likely to receive corporal punishment and more likely to receive ISS or "other." Students in Grades 3–8 were more likely to receive ISS and OSS and less likely to receive "other." There were also some more nuanced differences by infraction type. Tables for Grades 3–8 and Grades 9–12 are available from the authors by request.

[11]We conducted the Oster (2017) test using the psacalc user-written Stata command.

[12]Tables are available by request.

[13]For example, in the full sample, each CY ISS is associated with a 0.54 PPT lower likelihood of grade retention relative to OSS. On a base grade retention rate of 2.9%, this represents a 18.6% decline. In the ever-disciplined sample, each CY ISS is associated with a 0.8 PPT decrease in the likelihood of grade retention, relative to receiving OSS. On a base grade retention rate of 6.7%, this represents an 11.9% decline. In the ever-excluded sample, each CY ISS is associated with a 1.35 PPT decrease in the likelihood of grade retention, relative to receiving OSS. On a base grade retention rate of 11.6%, this represents an 11.6% decline. Results are available by request.

[14]We modeled exit from the dataset using a similar approach as in our main models and found that ALE and expulsion were associated with a 1.6 to 1.9 PPT increase in the likelihood of attrition from the data. OSS was not associated with a statistically significantly higher risk of attrition, except when compared to less exclusionary consequences for similar types of infractions.

[15]We estimated models that drop all observations for any students who were ever completely missing attendance data or for whom their days attended totaled less than 30 days in any given year, limiting the sample to students who attended school for at least one sixth of a typical school year. While many of the estimates on expulsion, ALE, no action/warning, and other were noisily estimated and lost significance in these new samples, the estimated coefficients on OSS, ISS, and corporal punishment were quite stable to this sample restriction.

[16]With the goal of addressing this, in earlier versions of this work, we estimated dynamic panel data models using within-student variation to identify the relationship between exclusionary discipline and academic achievement and found that accounting for student heterogeneity greatly attenuated the estimated relationships. However, given the assumptions required and challenges with the data available, there was still not strong support for causal identification in that case.

REFERENCES

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, *113*(1), 151–184.

Anderson, K. P. (2018). Inequitable compliance: Implementation failure of a statewide student discipline reform. *Peabody Journal of Education*, *93*(2), 244–263.

Anderson, K. P., & Ritter, G. W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a U.S. state. *Education Policy Analysis Archives*, *25*(49), 1–33.

Anyon, Y., Jenson, J. M., Altschul, I., Farrar, J., McQueen, J., Greer, E., Downing, B., & Simmons, J. (2014). The persistent effect of race and the promise of alternatives to suspension in school discipline outcomes. *Children and Youth Services Review*, *44*, 379–386.

Arcia, E. (2006). Achievement and enrollment status of suspended students: Outcomes in a large, multicultural school district. *Education and Urban Society*, *38*(3), 359–369.

Balfanz, R., Byrnes, V., & Fox, J. (2014). Sent home and put off-track: The antecedents, disproportionalities, and consequences of being suspended in the ninth grade. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *5*(2), Article 13.

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

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148.

Carpenter, D. M., & Ramirez, A. (2007). More than one gap: Dropout rate gaps between and among black, Hispanic, and white students. *Journal of Advanced Academics*, *19*(1), 32–64.

Choi, Y. (2007). Academic achievement and problem behaviors among Asian Pacific Islander American adolescents. *Journal of Youth Adolescence*, *36*(4), 403–415.

Chin, J. K., Dowdy, E., Jimerson, S. R., & Rime, W. J. (2012). Alternatives to suspensions: Rationale and recommendations. *Journal of School Violence*, *11*(2), 156–173.

Chu, E. M., & Ready, D. D. (2018). Exclusion and urban public high schools: Short-and long-term consequences of school suspensions. *American Journal of Education*, *124*(4), 479–509.

Cobb-Clark, D. A., Kassenboehmer, S. C., Le, T., McVicar, D., & Zhang, R. (2015). Is there an educational penalty for being suspended from school? *Education Economics*, *23*(4), 376–395.

Fabelo, T., Thompson, M. D., Plotkin, M., Carmichael, D., Marchbanks, M. P., & Booth, E. A. (2011). *Breaking schools' rules: A statewide study of how school discipline relates to students' success and juvenile justice involvement*. New York, NY: The Council of State Governments Justice Center & Public Policy Research Institute.

Fronius, T., Persson, H., Guckenberg, S., Hurley, N., & Petrosino, A. (2016). *Restorative justice in U.S. schools: A research review*. Retrieved from: http://jprc.wested.org/wp-content/uploads/2016/02/RJ_Literature-Review_20160217.pdf

Hashim, A. K., Strunk, K. O., & Dhaliwal, T. K. (2018). Justice for all? Suspension bans and restorative justice programs in the Los Angeles Unified School District. *Peabody Journal of Education*, *93*(2), 174–189.

Hinze-Pifer, R., & Sartain, L. (2018). Rethinking universal suspension for severe student behavior. *Peabody Journal of Education*, *93*(2), 228–243.

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions*, *11*(3), 133–144.

Kaufman, J. S., Jaser, S. S., Vaughan, E. L., Reynolds, J. S., Di Donato, J., Bernard, S. N., & Hernandez-Brereton, M. (2010). Patterns in office discipline referral data by grade, race/ethnicity, and gender. *Journal of Positive Behavior Interventions*, *12*, 44–54.

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

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

Morrison, G. M., Anthony, S., Storino, M. H., Cheng, J. J., Furlong, M. J., & Morrison, R. L. (2001). School expulsion as a process and an event: Before and after effects on children at risk for school discipline. *New Directions for Youth Development*, *92*, 45–71.

Nicholson-Crotty, S., Birchmeier, Z., & Valentine, D. (2009). Exploring the impact of school discipline on racial disproportion in the juvenile justice system. *Social Science Quarterly*, *90*(4), 1003–1018.

Noltemeyer, A. L., Ward, R. M., & Mcloughlin, C. (2015). Relationship between school suspension and student outcomes: A meta-analysis. *School Psychology Review*, *44*(2), 224–240.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*. doi: 10.1080/07350015.2016.1227711

Raffaele-Mendez, L. M. (2003). Predictors of suspension and negative school outcomes: A longitudinal investigation. *New Directions for Youth Development*, *99*, 17–33.

Sartain, L., Allensworth, E. M., & Porter, S. (with Levenstein, R., Johnson, D. W., Huynh, M. H., Anderson, E., Mader, N., & Steinberg, M. P.) (2015). *Suspending Chicago's students: Differences in discipline practices across schools*. Chicago: The University of Chicago Consortium on Chicago School Research.

Scott, T. M., & Barrett, S. B. (2004). Using staff and student time engaged in disciplinary procedures to evaluate the impact of school-wide PBS. *Journal of Positive Behavior Interventions*, *6*, 21–27.

Sharkey, J. D., & Fenning, P. A. (2012). Rationale for designing school contexts in support of proactive discipline. *Journal of School Violence*, *11*(2), 95–104.

Skiba, R. J., Arredondo, M. I., & Williams, N. T. (2014). More than a metaphor: The contribution of exclusionary discipline to a school-to-prison pipeline. *Equity & Excellence in Education*, *47*(4), 546–564.

Skiba. R. J., Chung, C., Trachok, M., Baker, T., Sheya, A., & Hughes, R. (2014). Parsing disciplinary disproportionality: Contributions of infraction, student, and school characteristics to out-of-school suspension and expulsion. *American Educational Research Journal*, *51*(4), 640–670.

Steinberg, M. P., & Lacoe, J. (2016). What do we know about school discipline reform? *Education Next*. Retrieved from: http://educationnext.org/what-do-we-know-about-school-discipline-reform-suspensions-expulsions/

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

Suh, S., & Suh, J. (2007). Rick factors and levels of risk for high school dropouts. *Professional School Counseling*, *10*(3), 297–306.

Swanson, E., Erickson, H. H., & Ritter, G. W. (2017). *Examining the impacts of middle school disciplinary policies on 9th grade retention* (EDRE Working Paper 2017–11). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952972

Vincent, C. G., & Tobin, T. J. (2011). The relationship between implementation of school-wide positive behavior support (SWPBS) and disciplinary exclusion of students from various ethnic backgrounds with and without disabilities. *Journal of Emotional and Behavioral Disorders*, *19*(4), 217–232.

Wolf, K. C., & Kupchik, A. (2017). School suspensions and adverse experiences in adulthood. *Justice Quarterly*, *34*(3), 407–430.

AUTHORS

KAITLIN P. ANDERSON, PhD, is a postdoctoral research associate at the Education Policy Innovation Collaborative at Michigan State University, 201F Erickson Hall, 620 Farm Lane, East Lansing, MI, 48824; *ande2018@msu.edu*. Her research focuses on student discipline, teacher quality, school choice, and educational access and equity.

GARY W. RITTER, PhD, is a professor and dean at the Saint Louis University School of Education, Fitzgerald Hall Suite 130, 3500 Lindell Boulevard, St. Louis, MO, 63103; *gary.ritter@slu.edu*. His current areas of research interest are student discipline policy, teacher quality, postsecondary access for low-income students, and the implementation and evaluation of programs aimed at improving educational outcomes for low-income students.

GEMA ZAMARRO, PhD, is an associate professor and 21st Century Endowed Chair in Teacher Quality at the Department of Education Reform at the University of Arkansas, 219-B Graduate Education Building, Fayetteville, AR, 72701; *gzamarro@uark.edu*. Her research focuses on the evaluation of education policies, measurement and development of character skills, and teacher quality.

Manuscript received April 3, 2018
Revisions received November 4, 2018;
February 2, 2019
Accepted March 8, 2019