The Effectiveness and Consequences of Exclusionary School Discipline

Chunxiao Li *

November 22, 2016

Abstract

Exclusionary school discipline techniques, such as out-of-school suspension, are often criticized for their inability to improve students' behavior, their adverse effects on students' achievement outcomes and their disproportionate use on minority students. Using comprehensive data on all North Carolina public school students, I find that harsher disciplinary rules (measured by higher out-of-school suspension likelihood) significantly deter students from committing first offenses, but that they are less effective (or ineffective) for repeat offenses. I also find that their adverse effects on offending students' achievement outcomes, such as end-of-grade test scores and high school dropout probability, are much smaller than the effects documented in the existing literature. In addition, I find that harsher disciplinary rules could significantly improve the academic achievement of middle school students with no offense record. The method I use to identify causal effects combines the instrumental variable method and a machine learning cluster method (k-means) to carefully address endogeneity and selection issues in a big data context. These findings suggest that current policy reform of exclusionary school discipline should carefully balance its benefits and costs for different student populations.

Keywords: Education; Out of school suspension; Racial bias

JEL Classification Numbers: I20, I21, I24, I28

*Department of Economics, University of North Carolina at Chapel Hill, *e-mail:chunxiao@live.unc.edu*. I thank Donna Gilleskie, Jane Fruehwirth, Valentin Verdier, David Guilkey, Helen Tauchen, Ju Hyun Kim for their advice, guidance and support through this project. I also thank Jonathan Williams, Brian Mcmanus, Stephane Bonhomme, Thibaut Lamadon and participants at the UNC applied microeconomics workshops for their valuable feedback. I acknowledge data access provided by the North Carolina Education Research Data Center and computing resources provided by UNC ITS Research Computing.

1 Introduction

In U.S. public schools, exclusionary school discipline techniques, such as out-of-school suspension or expulsion, are commonly used methods to address student misbehaviors – ranging from severe misconduct (e.g., assaults at school) to minor offenses (e.g., disruptive behavior in the classroom). In the 2013-2014 school year alone, 2.8 million of the 50 million public school students were suspended out-of-school at least once, and 130,000 students were expelled.[1] The high rate of out-of-school suspension reflects the consequences of policies such as "zero tolerance," which emphasize tough punishment, including social exclusion, as a primary response to crime or misconduct (Skiba and Knesting 2001; Losen and Skiba 2010). However, this disciplinary practice is widely criticized for its inability to improve students' misbehavior and its adverse consequences for suspended students and the broader school community.

Some existing literature suggests that suspension does little to discourage misbehavior and may, in fact, encourage it (Wettach et al. 2015; Skiba and Rauch 2015). The literature also finds that suspension lowers academic achievement and raises school dropout rates of offending students (Raffaele Mendez 2003; Arcia 2006; Lee et al. 2011; Skiba and Rauch 2015). Students who are suspended or expelled from school are more likely to be involved in the justice system; this relationship is often referred to as the "school to prison pipeline" (Wald and Losen 2003). A recent study shows that high rates of school suspensions actually harm math and reading scores for non-suspended students (Perry and Morris 2014). Concerns about these adverse effects are amplified by findings of disparities in school disciplinary practices, especially the disproportionate representation of students of color in school suspension rolls. In the 2013-2014 academic year, while black students represented 15.5 percent of the public school student population, they comprised 39 percent of student out-of-school suspensions.[2] These findings suggest that the disciplinary practice is harmful to students,

[1] Civil Rights Data Collection, 2016, U.S. Department of Education Office for Civil Rights.
[2] Civil Rights Data Collection, 2016, U.S. Department of Education Office for Civil Rights.

and is particularly harmful to minority groups. They have motivated the U.S. Department of Justice and Education to release a school discipline guidance package in 2014 to reform discipline policies and practices, several states to enact new legislation, and schools to consider ongoing school discipline policy reform.[3]

Despite concern expressed in the literature and in the popular press, the effectiveness and consequences of the exclusionary school discipline technique remain controversial. Several issues inherent in identifying "causal effects" in this context, such as endogeneity and selection problems, have not been fully addressed in the literature, which mostly uses descriptive statistics or regressions with limited sets of observables. For example, the negative correlation between out-of-school suspension and suspended students' achievement may reflect the following causal relationships. First, principals may be more likely to suspend frequently badly-behaved or less-engaged students than other students (even for the same type of offense), and the lower achievement outcomes of suspended students may be because they are "bad apples." Second, principals may be more likely to suspend students who commit more serious offenses, and a more serious offense experience may create (or reflect) students who are "bad apples," which eventually leads to lower achievement outcomes. These problems cannot be fully addressed using regressions with limited sets of observables because the engagement levels of students and the severity levels of offenses are typically not observed by researchers.[4] Furthermore, they cannot be fully addressed with student- or school-level fixed effects (or similar methodologies) because some of the unobservables, such as the severity of offenses or other life-changing circumstances of students, are time-varying factors.[5]

[3]See the following website for the guidance package: http://www.ojjdp.gov/enews/14juvjust/140109.html; http://www2.ed.gov/policy/gen/guid/school-discipline/index.html. There is new related legislation in several states, for example, California (AB 420, 2014) and Illinois (SB 100, 2015).

[4]Even if we observe the type of offense (e.g., fighting), we typically do not observe additional details about the offense (e.g., severity).

[5]A fixed effect approach uses differences in punishment across the same type offenses for the same student or in the same school to identify the effect. However, the fact that the same type offenses were punished differently for the same student or in the same school is likely to reflect that some unknown sources that determine the punishment have changed, such as the severity of offenses. There are also other limitations for applying the fixed effect approach in the context. For example, the student fixed effect is not applicable when the student's outcome, such as ACT score, is not observed repeatedly.

By carefully addressing the identification issues, this paper studies the causal effects of exclusionary school discipline on students' in-school behaviors and achievement outcomes, such as end-of-grade test scores, dropout probability and ACT scores.[6] With regard to students' in-school behaviors, I separately identify a "general deterrence effect" and a "specific deterrence effect" of the discipline. As punishments, these discipline techniques may serve as threats to students who intend to commit offenses and deter them from infractions. I refer to this mechanism as "general deterrence effect" following the economics of crime literature.[7] In addition, the experience of a specific punishment may serve as an effective "wake-up call" and decrease students' likelihood of re-offending in the future. I refer to this mechanism as "specific deterrence effect."[8]

I also examine the impact of exclusionary school discipline on students' achievement outcomes by separately identifying its effects on the achievement outcomes of offending students, students with no offense record, and all students.[9] An out-of-school suspension experience may directly lower offending students' achievement outcomes by, for example, reducing classroom learning time. However, if the discipline reduces subsequent misbehavior, then offending or would-be offending students may experience higher achievement. In addition, since exclusionary discipline removes offending students from the school, this social exclusion prevents (or incapacitates) offenders from committing additional offenses in school for a period of time. I follow the economics of crime literature in calling this an "incapacitation effect." Therefore, if students' misbehaviors are harmful to peers' achievement outcomes, then the positive "general and specific deterrence effects" and the "incapacitation effect" may reduce these "harmful spillover effects" and lead to better average achievement

[6]Specifically, I focus on the effects of out-of-school suspension (including expulsion) in this paper because it is the major controversy in policy making. Since it is also the most severe punishment in the scholastic setting, the effects of using it (or using it with higher likelihood) may also represent the effects of using more severe exclusionary school discipline rules.

[7]The exclusionary school discipline may encourage students to commit offenses instead; in this case, the "general deterrence effects" would be negative.

[8]The experience of out-of-school suspension, for example, may be a "bad lesson" to students and increase their likelihood of re-offending instead. In this case, the "specific deterrence effect" is negative.

[9]A student with no offense record means that she does not have an offense record in my sample period.

outcomes for students with no offense record or for all students.

This study uses linked administrative data with detailed misbehavior records for all North Carolina public school students from the 2008-2009 to 2014-2015 academic years. It identifies the theoretical effects using an empirical strategy that combines the instrumental variable (IV) method with a method that addresses the student unobserved heterogeneity problem in the big data context. The identification strategy exploits important features of the data; namely, students and principals are followed across academic years and across schools. The IV is a measure of out-of-school suspension propensities of principal teams, which is constructed by principal team members' out-of-school suspension decisions in other schools. It is assumed to affect students' misbehavior or achievement outcomes only through the principals' discipline decisions in the current school.

A complicating factor is that differences in principals' discipline decisions may induce students with different unobservables to commit offenses. This behavior causes a selection issue when estimating the effects of the discipline decisions on offending students' re-offending propensity or achievement outcomes. The selection issue might not be fully addressed by the IV method if the out-of-school suspension propensity of a principal team is anticipated by students. Because traditional approaches to address this issue are quite computationally expensive given the sample size, I propose an empirical framework that combines the IV method with recent approaches that address the latent heterogeneity problem. Following the literature (Lin and Ng 2012; Bonhomme and Manresa 2015; Bonhomme et al. 2016a), I model the student unobserved heterogeneity in a flexible yet parsimonious way, which allows it to vary across different groups (types) of students, across different types of misbehaviors and across different academic years. The solution involves application of the k-means clustering algorithm, which is widely used in machine learning and other related fields (Forgy, 1965; Steinley, 2006). After including school fixed effects, the empirical strategy identifies the causal effects by comparing outcomes of otherwise identical students in the same identified unobserved heterogeneity groups in the same schools but assigned to principal teams with

different out-of-school suspension propensities.

Only a small amount of literature that studies the effects of the discipline attempts to address the endogeneity and selection problems. The most relevant work is Kinsler (2013), which uses data on middle school students in the three largest school districts in North Carolina in the 2000-2001 academic year. The study found significant "general deterrence effects" of out-of-school suspension. It also suggests that out-of-school suspension has an overall positive influence on middle school students' end-of-grade test scores. The study attempts to address the endogeneity and selection issues by including permanent unobserved student heterogeneity while jointly estimating structure parameters of student behavior, end-of-grade test score production, and principals' punishment decisions. One limitation of the paper is that it assumes, conditional on the observed student characteristics and the permanent unobserved student heterogeneity, discipline punishments are exogenous; they are predetermined by principals with identical preferences by their forward-looking evaluation of different parties' welfare at the beginning of the academic year. With this assumption, along with some other assumptions, the study achieves identification without an exclusion restriction that affects principals' punishment decisions but does not affect students' misbehaviors.

Several issues indicate that such an exclusion restriction is important for recovering the causal effects. First, principals may change the severity of overall punishments in response to students' behaviors and the realized offense rate within a school/academic year.[10] This creates a problem that is called "reverse causality" for estimating the "general deterrence effects," i.e., observed high (low) rates of out-of-school suspension might be the consequences of, but not the reasons for, high (low) misbehavior rates. Second, as discussed before, the endogeneity and selection problem might stem from time-varying unobserved factors, such as severity of offenses, which are not fully addressed by controlling for students' permanent unobserved heterogeneity. These problems indicate that the explicit exclusion restriction

[10]In North Carolina, as in most U.S. states, local boards of education establish their own disciplinary policies following the broad principles in the state statutes. However, the disciplinary policies often list a broad range of punishments for each type of misbehavior and, thus, school principals or assistant principals may, at their discretion, determine appropriate punishment for each student infraction instance.

(the IV) used in this paper, which stems from heterogeneous preferences of principals (in contrast to the identical preferences assumption in Kinsler (2013)), is important to recover the causal effects because it addresses the problems of "reverse causality" and time-varying unobserved factors.

After carefully addressing the identification issues, I find the following results. First, I find statistically significant positive "general deterrence effects" of harsher discipline rules (measured by out-of-school suspension likelihood) on students' first offenses.[11] The effects are heterogeneous for different types of offenses and different student subpopulations. For example, while a 10 percentage point increase in the out-of-school suspension likelihood reduces the mean rate of first offenses for most categories or types of offenses by 7 to 40 percent, the effect is not statistically significant for some types of offenses, such as "fighting." In addition, I find that the effect is generally smaller for students' repeat offenses and not statistically significant for students who have already had an out-of-school suspension experience. Second, I find suggestive evidence that the "specific deterrence effect" is either small or not statistically significant.

Third, and contrary to the existing literature, I find that the effects of suspension experience on offending students' achievement outcomes are either small or not statistically significant. For example, an OLS regression with a limited set of controls suggests that getting suspended could decrease an offending student's end-of-grade math test score by 0.2 standard deviations, but my preferred estimate suggests that this effect is not statistically significant. Furthermore, while the OLS regression suggests that suspension experience increases an offending high school student's dropout probability by 18 percentage points, my preferred estimate suggests that it is not statistically significant.

Finally, I find that harsher disciplinary rules have statistically significant positive effects on end-of-grade math scores of students with no offense record.[12] For example, I find that

[11] Comparing the magnitude of the general deterrence effects between this article and Kinsler (2013) is not straightforward because it uses days of suspension to construct the discipline measure.

[12] This result is contrary to Perry and Morris (2014).

a 10 percentage point increase in the out-of-school suspension likelihood in a school could increase the end-of-grade math scores of middle school students with no offense record by about 0.02 standard deviations. I also find that they have an overall positive effect on all middle school white students' end-of-grade math scores. However, I find that the effects are not statistically significant for other student populations or other achievement outcomes.

The rest of the paper is organized as follows. In section 2, I describe the data for this research. In section 3, I discuss the empirical framework. Section 4 provides some estimation details. Section 5 offers results. Section 6 concludes.

2 Data Description

2.1 Data Sources

This study uses administrative data from North Carolina public schools provided by the North Carolina Education Research Data Center (NCERDC). The data were originally collected by the North Carolina Department of Public Instruction (NCDPI) and the National Center for Education Statistics (NCES). They include all North Carolina public school students' disciplinary infraction records, academic records and other administrative information. They also include teachers and other licensed personnel's information, and other school-level statistics.[13]

The students' disciplinary infraction records were collected by the NCDPI through each local education agency's (LEA's) superintendent's office (Director/Principal's office in the case of charter schools). They were firstly reported by the school disciplinary data coordinator, and the principal is ultimately responsible for the data elements. Due to state and federal statutes and state Board of Eduction policies, a record of offense incidents in-

[13]Additional information on perceptions of school environments is obtained from a survey (NCTWCS) of all teachers, principals and other licensed personnel in North Carolina public and charter schools conducted by the North Carolina Professional Teaching Standards Commission (NCPTSC) and the Governor's office. Additional information on Positive Behavior Intervention and Support (PBIS) school recognition is collected from NCDPI website.

volving the following must be reported: 1.) any of the 17 criminal acts committed on a school campus or in connection with a school function;[14] 2.) any act resulting in an out-of-school suspension or expulsion; 3.) any in-school suspension received by an exceptional student;[15] 4.) any of the following acts, regardless of consequences assigned: fighting (or affray), bullying, discrimination, harassment, a violent assault not resulting in serious injury, communicating threats, gang activity, extortion, property damage, and possession or use of tobacco products. In addition to the offense incidents required to be reported by statutes and policies, other routine disciplinary incidents were also recorded and reported for analysis or administrative purpose. Guidelines have been created to ensure consistent reporting.

The academic records were collected by the NCDPI, and include end-of-grade test scores, end-of-course test scores, ACT scores and other academic record information of all North Carolina public school students. In North Carolina, students in grades 3-8 are required to take end-of-grade tests in reading and math, and students in grades 9-12 are required to take end-of-course tests for Algebra I and English 1 when they are enrolled for credit in these courses.[16] In addition, beginning in 2012-2013 academic year, every 11th grader is required to take the ACT college entrance exam as part of the new North Carolina Standard Course of Study. These test scores are used as academic achievement measures in this paper.[17]

The disciplinary infraction records were matched with the academic records and all other

[14] 17 reportable acts are: homicide, assault resulting in serious bodily injury, assault involving the use of a weapon, rape, sexual offense, sexual assault, kidnapping, robbery with a dangerous weapon, robbery without a dangerous weapon, taking incident liberties with a minor, assault on school personnel, bomb threat, burning of a school building, possession of alcoholic beverage, possession of controlled substance in violation of law, possession of a firearm or powerful explosive, and possession of a weapon. Robbery without a dangerous weapon was removed from this category to category 5 since 2010-2011 academic year. 17 (or 16) criminal acts are required to be reported to law enforcement before 2011-2012 academic year. Since 2011-2012 academic year, possession of alcoholic beverage, bomb threat, and burning of a school building are no longer required to be reported to police.

[15] Guidelines in later years (e.g., 2014/2015) require reporting any act resulting in-school suspensions. I don't find evidence that exceptional students are more likely to be reported than other students for in-school suspension in the data used for this project.

[16] An End-of-Course test in English 2 instead of English 1 has been required since the 2012-2013 academic year. The data also include records of students who were absent or exempt from the tests.

[17] Other academic records provided by NCERDC, such as end-of-grade or end-of-course test scores for writing, computer skills, biology, Algebra 2, Civics and Economics, U.S. History, Chemistry and Physics, and SAT scores are not used in this paper because they are not available for all the academic years (or only available for part of the students).

information by NCERDC.[18] Each student and licensed school personnel was assigned an unique identifier (a randomized number). This identifier allows me to follow a student or school personnel (e.g., principals or assistant principals) over time and across schools, which is important for my identification strategy.

2.2 Sample Construction

The disciplinary infraction data provided by NCERDC span academic years from 2000-2001 to 2014-2015. Beginning in the 2007-2008 academic year, reporting requirements for offenses have been greater, and the matching rate of infraction data with other data has largely increased.[19] In addition, in the 2013-2014 academic year, there may be a data imputing problem caused by an upgrade of the data system of the N.C. public schools.[20] Therefore, for most of the empirical work, I use data from the 2007-2008 to 2012-2013 academic years. I use the 2013-2014 and 2014-2015 academic year data mostly for robustness checks, except that the ACT score information and dropout information in these academic years are used for the main estimations.[21]

The data for charter schools are not used in this paper because the offense reporting requirement and rates can be different and school administrator information is largely unavailable. In addition, I use student observations of grades 3-12 since several common explanatory variables, such as economically disadvantaged status and limited English proficiency status, are not available for grades K-2.

In the data, there are 1,909,831 distinct public school students in grades 3-12 from the 2007-2008 to 2012-2013 academic years, who contribute 6,559,362 student-year obser-

[18]The detail of the matching process is on the website of NCERDC: http://childandfamilypolicy.duke.edu/research/nc-education-data-center/list-files-variables/. Students who were not matched are not used for this project.

[19]The matching rate is greater than 99 percent since the 2007-2008 academic year.

[20]Since the 2013-2014 academic year, N.C. public schools have upgraded from an NC WISE to a Pearson's PowerSchool product to report suspension data. There were reports that the new system experienced crashes and technique issues.

[21]The IV is constructed by using punishment records from 2007-2008 to 2014-2015 academic year. 2013-2014 and 2014-2015 academic year data are used to ensure enough observations to construct the IV.

vations.[22] For most of the empirical work, lagged student or school offense statistics are used as explanatory variables; in these cases, the 2007-2008 academic year data are not directly used in estimation. Further, I drop schools missing one or more years of data. These deletions result in 1,687,330 distinct students and 5,271,039 student-year observations. I use this sample to calculate the summary statistics in the next subsection. Depending on the empirical work, there may be additional data selections. For example, for estimating the effects of suspension on end-of-grade test scores, I use students in grades 3-8. For estimating the effects on ACT scores, I mainly use disciplinary data of students in grade 9-10 from the 2009-2010 to 2012-2013 academic years, with their ACT scores from the 2012-2013 to 2014-2015 academic years. For the effects on dropout probability, I mainly use grade 9 students' disciplinary data with the information that they finally graduate or dropout from high school.[23]

2.3 Descriptive Statistics and Data Issues

There are 3,951,754 recorded offense instances in the constructed sample, which were committed by 651,040 distinct students with 1,236,497 student-year observations. That means about 23 percent of student-year observations have at least one offense in the academic year, and about 38 percent of distinct students have at least one offense record in the sample period. In addition, about 56 percent of offending students re-offended in the same academic year.[24] About 20 percent of distinct students (349,611 distinct students) had at least one

[22] A student-year observation is calculated as one yearly observation per student if the student is ever enrolled and assigned an identifier in the data. The student might have several offenses or transfer to different schools within one year, but they are all regarded as one observation in this calculation. Students who are not matched between disciplinary data and other administrative data are not in this calculation and not used for this paper.

[23] Only first time grade 9 students (no students who repeat grade 9) are included in the sample. As discussed later, I use grade 9 students to avoid the dynamic selection problem. I drop grade 9 students in 2012-2013 academic year because I only observe the dropout information until 2014-2015 academic year. Since there might still be censored information for students who repeat grades, I do robustness checks by only using students from the 2008-2009 to 2010-2011 academic years.

[24] The calculation is based on student-year observations but not distinct students.

out-of-school suspension record in the sample period.[25]

The NCDPI classified offenses using about 90 offense types.[26] There were nearly 40 consequence types that were assigned to the offenses. The most commonly used consequence types were out-of-school suspension and in-school suspension, each representing about 30 percent of all consequences in the data. Most of the out-of-school suspensions were short-term ($\leq$ 10 days); only 1 percent of out-of-school suspensions were long-term ($>$ 10 days).[27] There were only 146 cases of expulsions in the sample. Most other consequence types were less severe punishments than suspensions, such as lunch detention or a warning.[28] My empirical work simplifies the punishment as either out-of-school suspension (including expulsion) or not, where the "not" category includes in-school suspension or other less severe punishments, in order to focus on the most severe exclusionary discipline.

To simplify the discussion, I provide two classifications of the offense types based on their similarities.[29] The first classification contains six categories - "violence," "drug," "disrespect," "truancy," "property" and "other offenses." This classification is used to discuss deterrence effects by category. A more detailed classification further divides the "violence" and "other" categories into four sub-categories respectively, resulting in twelve categories. This classification is used for constructing explanatory variables and identifying student unobserved heterogeneity. Appendix A shows the offense types in each category by classification. One concern for estimating the deterrence effects of out-of-school suspension is that additional punishments assigned by the juvenile justice system for criminal acts are not observed. Very few violent offenses (less than 4 percent) and property offenses (less than 0.5 percent) re-

[25]About 10 percent of student-year observations (577,886 student-year observations) were out-of-school suspended at least once in the academic year.

[26]Each offense instance may be described by multiple offense types. Since only about 1 percent of offenses have associated with them more than one offense types, I use the first (typically most serious) offense type to represent the offense. The number of offense types and the definition of each type changed across years. But they only changed slightly from 2007 to 2015 and most of the changes only involve adding new types.

[27]According to the North Carolina state statute, a long-term suspension must be assigned by the superintendent under a principal's recommendation (115C-390.7).

[28]There were 1 percent of offenses that were assigned the consequence of "alternative learning program." Of these, about 40 percent were assigned out-of-school suspension at the same time. This paper ignores "alternative learning program" as a type of punishment.

[29]The classifications also take into account the number of observations.

quired reporting to law enforcement. Most of the offenses in the "drug" category, however, required such reporting and might have resulted in additional punishment; results from this specification should be interpreted with caution.

Table 1 provides summary statistics for different categories of offenses and several types of offenses within categories.[30] The percentages of offenses punished by out-of-school suspension (Column 3-4) show that out-of-school suspension was frequently used even for minor offenses, such as disruptive behavior and excessive tardiness.[31] The generally higher rates of out-of-school suspension for students' second offenses than for their first offenses reflect that the escalating punishment rules were commonly used in education practices. It motivates my empirical framework to separately use the first and second punishment to study the deterrence effects.

According to the data reporting requirement, almost all types of offenses in the "violence," "drug" or "property" categories must be reported regardless of the consequences assigned. However, one concern is whether other minor offenses, such as from the category "truancy," are well reported. Column 5 lists the percents of schools with at least one reported offense in the academic year, which provides some evidence of this reporting requirement. The percentages suggest that, in addition to "violence" offenses, "disrespect" was also widely reported in many schools.[32] Furthermore, although schools were required to report most of the "drug," and "property" offenses, there was a relative low percentage of schools that actually reported them. This low percentage is likely due to lower offense or catching rates. Note that some less severe offenses, such as disruptive behavior, are also widely reported. But for other less severe offenses, such as truancy and excessive tardiness, the percentages are relative low. A further check by different school levels shows that the reporting rate of these offenses changes largely by school levels. More than 84 percent of middle school or high

[30] "Excessive tardiness" and "disruptive behavior" are from "other" category, which account for more than 53 percent of offense cases of the category.

[31] The percentages also show that the sample includes a large portion of offenses that were assigned less severe punishments than out-of-school suspension.

[32] Since, intuitively, likelihood of being caught committing these offenses should be high, I use these offenses to construct my main instrumental variable.

school school-year observations had reported at least one "truancy" case, and more than 50 percent of middle school or high school school-year observations had reported at least one excessive tardiness case. However, it is likely that many minor offenses were not reported. Therefore, in addition to the econometric effort I will make to address this issue, results from these offenses should be interpreted with caution.

To further illustrate the data, Table 2 separately reports sample means of student characteristics for the student-year observations with and without any offense record in the academic year. For the observations with any offense record, I further divide them into those who were not punished by out-of-school suspension in the academic year and those who were punished by out-of-school suspension ever in the academic year. The table shows that the three groups differ considerably along all the dimensions, which indicates the selection problems of the students into the offender group and into the out-of-school suspended group. While the ratios of white students decrease from the no offense group to the offender group, and from the not out-of-school suspended offender group to the out-of-school suspended group, the ratios of black students differ in these groups in an opposite direction. Black student-year observations account for 23 percent of the no offense group, 33 percent of the not out-of-school suspended offender group, but account for 51 percent of the out-of-school suspended group, which indicates the over-representation of black students in out-of-school suspension rolls. Female students make up a smaller ratio in the offender group and a further smaller ratio in the out-of-school suspended group. Economically disadvantaged students account for a higher ratio in the offender group and a further higher ratio in the out-of-school suspended group.[33] The offender group or the out-of-school suspended group were more likely to be physically or intellectually disabled, less likely to be academically and intellectually gifted, and more likely to be above typical age in the grade or be repeating the grade in the academic year than the no offense group. They also had worse lagged test scores. Grade

[33]Economically disadvantaged students are the students receiving free or reduced price meals. The eligibility for free or reduced lunch is determined by family size and family income. The most recent criteria could be found on http://www.dpi.state.nc.us/newsroom/news/2015-16/20150814-01

9 was the grade with the highest offense ratio and out-of-school suspension ratio.

3 Empirical Framework

In this section, I introduce the empirical framework used to estimate the "general deterrence effects," the "specific deterrence effects," and the effects of exclusionary school discipline on students' achievement outcomes. The theoretical motivation for the empirical model is provided in a detailed online Appendix. After detailing the main equations, I discuss econometric issues that must be addressed in order to obtain unbiased causal effects.

3.1 Main Equations

Let $D_{1is\tau}$ ($D_{2is\tau}$) indicate whether or not student i in school s committed a first (or second) offense in academic year τ. Let Y_{is} be the student's achievement outcome (i.e., end-of-grade test score, ACT score or an indicator of high school dropout or graduate). I define the following empirical models:

$$D_{1is\tau} = \beta_{10} + \beta_{11} X^a_{is\tau} + \alpha_{11} P^*_{1is\tau} + \phi^{sch}_{1s} + \epsilon_{1is\tau} \tag{1}$$

$$D_{2is\tau} = \beta_{20} + \beta_{21} X^b_{is\tau} + \alpha_{21} P^*_{2is\tau} + \alpha_{22} P_{1is\tau} + \phi^{sch}_{2s} + \epsilon_{2is\tau} \quad \text{if} \quad D_{1is\tau} = 1 \tag{2}$$

$$Y_{is} = \beta_{30} + \beta_{31} X^c_{is\tau} + \alpha_{31} P_{is\tau} + \phi^{sch}_{3s} + \epsilon_{3is\tau} \quad \text{if} \quad D_{1is\tau} = 1 \tag{3}$$

$$Y_{is} = \beta_{40} + \beta_{41} X^d_{is\tau} + \alpha_{41} P^*_{is\tau} + \phi^{sch}_{4s} + \epsilon_{4is\tau} \tag{4}$$

where $X_{is\tau}$ is a vector of control variables to be defined later; $P^*_{1is\tau}$ ($P^*_{2is\tau}$) is the potential punishment for the student if she commits the first (second) offense; $P_{1is\tau}$ is the actual punishment (out-of-school suspension or not) received for her first offense (i.e., if $D_{1is\tau} = 1$). In estimation, $P_{is\tau}$ is an indicator of whether the student was ever suspended out-of-school in academic year τ; in a robustness check, $P_{is\tau}$ is an indicator of whether the student was

suspended out-of-school for her first offense in academic year τ.[34] $P^*_{is\tau}$ measures the potential punishment rules for all her offenses. ϕ^{sch}_s represents school fixed effects and $\epsilon_{is\tau}$ is the error term.

The coefficients α_{11} and α_{21} capture the "general deterrence effects," which describe the effects that result from the threat of a punishment for misbehavior. The coefficient α_{22} captures the "specific deterrence effect," which describes the role of a previous punishment (as one's punishment experience) on re-offending. The first equation captures the general deterrence effect of out-of-school suspension on "ever misbehaving or not" in the academic year, and the second equation isolates the general and specific deterrence effects on "recidivism or not" in the academic year. For these equations, I also separately estimate the effects by categories (or types) of misbehavior, where for m type misbehavior, for example, the outcomes are defined as "ever committing m type misbehavior or not in the academic year," and "re-offending m type misbehavior or not in the academic year."[35]

The coefficient α_{31} captures the effects of own out-of-school suspension experience on offending students' achievement outcomes. I separately estimate equation (4) using all students or only students with no offense record in the sample period; the coefficient α_{41} captures the effects of exclusionary school discipline on the achievement outcomes of all students or students with no offense record respectively.

3.2 Estimation Concerns

Several issues must be addressed in order to recover unbiased causal effects of the exclusionary school discipline on students' behavior and achievement outcomes.

[34] If student i has transfer experience within academic year τ, $P_{1is\tau}$ and $P_{is\tau}$ are defined by actual punishments received for her offenses regardless of in which school they were assigned. Thus, more precise notations for $P_{1is\tau}$ and $P_{is\tau}$ for transfer students should be $P_{1i\tau}$ and $P_{i\tau}$.

[35] I also show results from separately estimating the model for different sub-populations, such as white or black students. As a robustness check, I also estimate a specification that defines the outcome variables as the student's first (second) offense in middle school or high school, instead of the first (second) offense in an academic year.

Unobserved Potential Punishments

In estimating the general deterrence effects, α_{11} and α_{21}, potential punishments for those who didn't misbehave if they had misbehaved (or those who misbehaved but were not caught or reported by the administrators) are not observed by the researcher. Since the data record different punishments for the same type of misbehavior in the same school in the same academic year, a potential punishment should not be assumed. The problem is similar for estimating α_{41}. To address this problem, I use proxy variables, $\hat{P}_{1s\tau}$, $\hat{P}_{2s\tau}$ and $\hat{P}_{s\tau}$, to approximate $P^*_{1is\tau}$, $P^*_{2is\tau}$ and $P^*_{is\tau}$.[36] For misbehavior category c, for example, $\hat{P}_{1s\tau}$ ($\hat{P}_{2s\tau}$) is defined by a "normalized" (by offense types) rate of out-of-school suspension assigned to the first offense (the second offense) of c category misbehavior of all the students in school s in academic year τ.[37] $\hat{P}_{s\tau}$ is defined by a "normalized" rate of out-of-school suspension assigned to all offenses in school s in academic year τ. I refer to these proxies as "Disciplinary Punishment Indexes (DPI)" of the school. I discuss the details of their construction in section 4. Since $\hat{P}_{1s\tau}$ ($\hat{P}_{2s\tau}$, $\hat{P}_{s\tau}$) is the same for all students in the same school in the same academic year, it approximates the actual punishments for each would-be offender with measurement error. That is,

$$P^*_{1is\tau} = \hat{P}_{1s\tau} + e_{1is\tau} \tag{5}$$

$$P^*_{2is\tau} = \hat{P}_{2s\tau} + e_{2is\tau} \tag{6}$$

$$P^*_{is\tau} = \hat{P}_{s\tau} + e_{is\tau} \tag{7}$$

[36] I do not define the general deterrence effects by students' beliefs or information, which are mechanisms that the general deterrence effects work through. That is, there might be no general deterrence effects because students don't know the disciplinary rules or they have incorrect beliefs on the rules.

[37] I use current year (instead of past year) punishment decisions to construct the proxy variables because punishments assigned for the current year offending students should exactly be the potential punishments they would get when they make offense decisions, and these punishments should also better reflect the current year disciplinary policy for other students. The idea of not using past year data to construct the measure also follows the spirit of "Lucas Critique." The problem in this context could be, for example, if the ineffectiveness of past year lenient punishments had led the principals to change to a harsher discipline rule in new academic year, which was announced to (or expected by) students and changed their behavior, using past punishments would lead to a false conclusion that the lenient discipline changed students' behavior.

Note that the proxy variables are only used for estimating α_{11}, α_{21} and α_{41}. For estimating the specific deterrence effect and the effects on offending students' outcomes, α_{22} and α_{31}, the actual "out-of-school suspension or not" indicator, $P_{1is\tau}$ and $P_{is\tau}$, are used because they are observed.[38]

Observed Heterogeneity

An administrator's punishment decision for an offense of a student is often correlated with the observed characteristics of the offense, the student, and the school. These characteristics might affect the student's misbehavior decision, re-offending decision, or her achievement outcomes as well.

To address this confounding bias, I include a comprehensive set of control variables. The control variable vectors, $X^a_{is\tau}$, $X^b_{is\tau}$, $X^c_{is\tau}$, and $X^d_{is\tau}$, each include the student's observables and her lagged math and reading scores (Table 2) and characteristics of her peers in the same grade.[39] They also all include time-varying school observables (Table 3), such as school-level offense and punishment statistics in the previous academic year, and teacher and school quality measures in the current academic year. Further, $X^a_{is\tau}$, $X^b_{is\tau}$, and $X^d_{is\tau}$ include her offense frequencies of each misbehavior category (twelve categories) in the previous academic year.[40] The vector $X^c_{is\tau}$ includes her offense frequencies of each misbehavior category in the current academic year; in a robustness check, it includes her offense frequencies of each misbehavior category in the previous academic year. Therefore, in the main estimation, α_{31} does not include the potential effect of exclusionary school discipline on achievement

[38]Robustness checks are done by using the proxy variables to estimate the specific deterrence effects and the effects on offending students' outcomes. The results are consistent.

[39]Same grade peers' characteristics include the ratios of black students (to the whole student population), other minority students, female students, exceptional students, AIG students, students with limited English proficiency, students who repeated grades this academic year, students who are above the typical age in the grade, and economically disadvantaged students, and the means of peers' last year math standard scores and peers' last year reading standard scores.

[40]In addition, $X^b_{is\tau}$ includes the specific type of her first misbehavior in the current academic year. For estimation by category of misbehavior, I include the number of her offenses in other categories before the first offense in the given category. The control variables that describe a student's offense history may be endogenous because they are correlated with the student's unobserved heterogeneity. I present how I address the student's unobserved heterogeneity problem later in this section.

outcomes from the offending students' behavior changes in the current academic year (i.e. does not include the indirect effect that works through the "specific deterrence effect"); in the robustness check, α_{31} includes this potential indirect effect.[41] Moreover, for estimation by a specific category of misbehavior, $X^a_{is\tau}$ and $X^b_{is\tau}$ include her same grade peers' offense rates of other offense categories in the current academic year.[42] The vector $X^c_{is\tau}$ includes same grade peers' offense rates in all offense categories in the academic year.

For equation (2), in addition to the control variables described above, I add the variable, "remaining number of in-school days" for the student in the academic year, which is calculated by subtracting the days of out-of-school suspension for the first punishment from the total remaining days in the academic year after her first offense. Inclusion of this variable addresses a potential concern when estimating the specific deterrence effect. That is, if a student was out-of-school suspended (or expelled) for the first offense, then she would not be able to re-offend (in-school) during her out-of-school suspension. The "incapacitation effect" might confound the "specific deterrence effect" because a lower re-offending likelihood may be due to less available time for another offense during the academic year and not due to the "wake-up call" effect of the suspension. Using the additional control variable, I am able to separate the "specific deterrence effect" from the "incapacitation effect" by comparing students who have the same remaining number of days to commit the second offense.[43]

[41] In the robustness check, $X^c_{is\tau}$ includes the specific type of her first misbehavior in the academic year.

[42] I carefully choose to not control for the peers' offense rates of the same offense category model to let the general deterrence effect for violent behavior, for example, include both its direct effects on a students' violent behavior, and its indirect effects by affecting the student's peers' violent behavior. I control for offense rates of other categories because one concern for my IV estimates (introduced later) by category of offense is that the IV may change the discipline decisions for other categories as well. This response might be another channel that the IV affects the misbehavior outcomes in the discussed category if the discipline rule for the category is not representative for other categories. Robustness checks are done with controls for peers' offense rates in all offense categories to achieve the direct effects (not through peer effects) of the discipline. Robustness checks are also done without controlling for any peers' misbehavior rate to address the concern that controlling for the peers' offense rate in other categories may cause an additional endogeneity problem for estimating the deterrence effects.

[43] The remaining days variable might be endogenous because it is correlated with the student's unobserved heterogeneity. I present how I address the student's unobserved heterogeneity problem later in this section.

Unobserved Factors

Despite the inclusion of the comprehensive set of observables, the estimation issues may not be fully addressed because an administrator's punishment decision for an offense of a student might depend on the unobserved characteristics of the offense, the student, and the school. A related problem is, student misbehaviors are not observed if they were not caught or reported. The unobserved misbehaviors would be incorrectly regarded as no misbehavior, for example, for the outcome variables in equations (1) and (2). Therefore, the unobserved catching or reporting likelihoods lead to biased estimated effects of discipline if they are correlated with the punishment variables, $\hat{P}_{1s\tau}$, $\hat{P}_{2s\tau}$ or $P_{1is\tau}$. To address these estimation issues, first, I include school fixed effects, ϕ_s^{sch}, to control for the unobserved factors or unobserved detecting or reporting likelihoods that are different across schools. The biases due to unobserved detecting and reporting rates, for example, are addressed if the rates are constant or are not correlated with the punishments in the same school. However, there is still a concern about other unobserved factors, such as the unobserved severity of offenses, which may vary within schools. I use an instrumental variable (IV) to address part of the problem. The IV also addresses a "reverse causality" problem when estimating the general deterrence effects, which suggests that high offense rates may be the reason for but not the consequence of more severe punishments in a school. The IV is also helpful for addressing the unobserved reporting and catching probabilities, as I only need to assume that they are not correlated with the IV in the same school.

The IV describes the out-of-school suspension propensity of a principal team. The idea of constructing the IV by using the "punishment propensity" of people is similar in spirit to Kling (2006) and Aizer and Doyle (2015) for the identification of causal effects of incarceration, and Doyle (2007) for the identification of causal effects of foster care. The IV is constructed as the "normalized" out-of-school suspension rates in other schools in which the team members worked. Construction of the IV is detailed in section 4.1. To understand how the IV works, consider a simplified example with two types of administrator teams:

lenient and tough. The tough team prefers to use out-of-school suspension more than the lenient team even for the same type of offense committed by identical students in the same school. The differences in re-offending rates and achievement outcomes between these two teams are used to identify local average treatment effects (LATE) of the specific deterrence on re-offending or of suspension on offending students' achievement outcomes (Imbens and Angrist 1994). The LATE measures the effects of punishments on students who wouldn't be punished by the lenient team but would be punished by the tough team. For example, suppose the severity of the offenses is the main determinant of punishment decisions; let severity be measured between one and ten, with ten as the most severe misbehavior. If the tough team out-of-school suspended all the students who committed offenses with severity more than three and the lenient team out-of-school suspended all the students who committed offenses with severity more than seven, the LATE measures the effects of past punishments on the re-offending rates or achievement outcomes for the "compliers" – the students who committed offenses with severity between three and seven. Since my actual IV variable is a continuous variable, my IV strategy identifies a weighted average of the compliers induced by each marginal change in the IV values (Heckman and Vytlacil 2005). I can use the identified effects to extrapolate the effects on other sub-populations only under the assumption that the effects are homogeneous for them. An important condition for the identification of the LATE is monotonicity. It suggests that any student who is out-of-school suspended by the lenient team would also be out-of-school suspended by the tough team, and any student who is not out-of-school suspended by the tough team would not be out-of-school suspended by the lenient team. For my continuous IV, the monotonicity should be satisfied for every marginal change of the IV. While the assumption cannot be fully tested, I show some indirect evidence in the next section to check the assumption.

Selection

While the IV can address some concerns about unobserved influences, an additional concern is selection of students (in terms of their unobservables) into the offending group because of punishment differences. For example, if students who were deterred from committing first offenses by out-of-school suspension are different from those who were not deterred in unobserved ways, then the population of offending students for the estimation of equation (2) or (3) may differ in unobserved ways depending on whether or not they would be punished by out-of-school suspension. Therefore, the identification of the effects of suspension on offending students' future outcomes might be confounded by unobserved determinants, which differ for suspended and not suspended students. This selection issue may not be fully addressed by the IV method because for any IV, if the potential punishment differences generated by the IV are anticipated by students, they may selectively (in terms of their unobservables) enter into the offending group due to these differences. In such a case, the IV would be correlated with the unobserved factors.[44]

To illustrate the problem, suppose that there is another unobserved student characteristic, such as "sensitivity to punishments," which reflects a student's level of self-control, or school engagement or other unobserved characteristic. Some students may be not sensitive to the out-of-school suspension; they are equally likely to misbehave (or not misbehave) regardless of whether they would or would not be out-of-school suspended. There may be other students who are sensitive to the suspension; they are more likely to misbehave if the misbehavior would not be punished by out-of-school suspension, and are less likely to misbehave if the misbehavior would be punished by out-of-school suspension. I call this latter group the "punishment compliers." Among the offenses that the tough team would punish but the lenient team would not punish, the offending students under the tough team are more likely to be the non-sensitive students than those under the lenient team because the latter team would

[44] An instrumental variable may address this problem if the discipline differences generated by it are a "surprise" for students.

get more "punishment compliers." For estimating equation (2) and (3), the IV could not fully address the identification issues when the sensitivity of students (or the student unobserved characteristics that it reflects) are correlated with outcomes in these equations. In addition, $\hat{P}_{1s\tau}$, $\hat{P}_{2s\tau}$, and $\hat{P}_{s\tau}$ are constructed by the punishments for the offending students, who might be from selected groups according to the above argument. Therefore, the measurement error of these variables might be correlated with the student's sensitivity level and thus be correlated with the IV; the IV method might also not fully address the problem for estimating α_{11}, α_{21}, and α_{41}.

To address the problem, I propose an empirical strategy that combines the IV method with explicit controls for student unobserved heterogeneity. Conditional on this unobserved student heterogeneity (e.g., the unobserved sensitivity level), the IV estimation would only compare the achievement (or behavior) outcomes of students with the same sensitivity levels who are assigned to different principal teams. The information I use for identifying the unobserved heterogeneity is students' misbehavior decisions for all types of offenses across all academic years conditional on observables and school fixed effects. I assume that the student unobserved type is reflected by her misbehavior decisions across time and across different types of misbehaviors. Conditional on all observed factors, students who are non-sensitive to the punishments, for example, might frequently (or never) commit offenses across different types of misbehaviors or across time although the punishment rules change across these dimensions.

Following recent approaches that address the latent heterogeneity problem (Lin and Ng 2012; Bonhomme and Manresa 2015; Bonhomme et al. 2016a), I model the unobserved heterogeneity in a flexible yet parsimonious way. Specifically, I assume that there are a relatively small number of distinct groups (types) of students, and students within the same group share same misbehavior type patterns and time patterns of unobserved heterogeneity. Group membership is estimated from the data, and is assumed to not change across misbehavior categories (or types) or across time. The heterogeneity is allowed to change

across different groups (types) of students, across different types of misbehaviors, and across time. Using $n = 1, 2, 3, 4$ to represent equation (1) through (4), the error term, $\epsilon_{nis\tau}$, can be decomposed as follows:

$$\epsilon_{nis\tau} = \theta_{ng_is\tau} + v_{nis\tau} \tag{8}$$

The first component, $\theta_{ng_is\tau}$, represents student i's unobserved heterogeneity in academic year τ. The term is also different for different categories (or types) of misbehaviors.[45] The subscript $g_i \in \{1, \ldots, G\}$ represents the student's group membership. The second component, $v_{nis\tau}$, is an error term that is not correlated with the IV.

Estimation of the student unobserved heterogeneity proceeds in two steps. The first step is to partition all students into G groups. It is solved by an application of the k-means clustering algorithm, which is widely used in machine learning and other related fields (Forgy 1965; Steinley 2006). In the second step, the group-specific unobserved heterogeneity is estimated separately in equations (1) though (4) with all other parts of the models by imputing the estimated group memberships of students into the models.

The approach that solves the unobserved heterogeneity problem is related to, but different from, finite mixture models or other "random effect" estimation models, which rely on assumptions that restrict the correlation between unobserved heterogeneity and the covariates. In contrast, it is in close analogy with fixed effects, which leaves the correlation between the unobserved heterogeneity and other covariates unrestricted (Bonhomme and Manresa 2015; Bonhomme et al. 2016a). Note that student unobserved heterogeneity is likely to be correlated with the school fixed effects or other school observables due to the non-random sorting of students into schools. It is also likely to be correlated with other student-level explanatory variables, such as the student's past offense history, past test scores, "the remaining days variable," and the peer variables. A "random effect" style approach to solve these problems would require modeling all these mechanisms and jointly estimating the structure equations with additional exclusion restrictions; the computation would be infeasible given our sample

[45] Misbehavior category or type subscripts are suppressed for simplicity.

size and the identification might have to rely on strong assumptions. Furthermore, traditional student fixed effects do not work in this context for the second and third equation because they are estimated only for offending students. Students who do not repeatedly commit offenses across years have no counterfactuals to cancel out the fixed effects; and directly adding individual indicators would result in the incidental parameters problem.[46] They may not work with the first and fourth equations either. First, these equations include lagged outcomes as explanatory variables. Second, when the school fixed effects are also included, the transfer of students across schools causes a two-way fixed effect problem.

The proposed approach also effectively uses the information on students' numbers of offenses of different types, and offers a flexibility that allows the unobserved heterogeneity to change across type of offenses and across time. This flexibility also relaxes the assumption for the validity of the IV. The details of estimation of the student unobserved heterogeneity are provided in section 4.3.[47]

4 Estimation Details

4.1 Instrument Construction and Validity

The IV describes the out-of-school suspension propensity of a principal team. Its construction exploits an important data feature – principals or assistant principals transferring across schools over time. For each principal team (principals and assistant principals) in each school in each academic year, if there were members who had principal experience in other schools, I calculate a "normalized" out-of-school suspension rate for these schools. I use it to construct the IV for suspension decisions in the school of concern.[48] The underly-

[46]The approach used in this paper solves these problems by reducing the dimensionality of the student dummies, and build counterfactuals by the same group students.

[47]In addition to the outlined econometric concerns, selection plagues estimation of the discipline effects in equations (3) and (4) when the outcomes are dropout indicator or ACT scores. I discuss this issue in section 5 when presenting the results. I also discuss other limitations in section 6.

[48]The IV is defined at the school level but not at the principal level because for each case I don't observe which principal or assistant principal assigned the consequence.

ing assumption is that principals who were more likely to use out-of-school suspensions in other schools might be more likely to use them in the current school because it may reflect their preferences or beliefs about the effectiveness of out-of-school suspension (or harsher discipline rules). I do not include any out-of-school suspension decisions (for any types of offenses in any academic years) in the current school in construction of the IV because they may have direct effects on the students' outcomes in the current school, which may violate the IV assumption.

For student i in school s in academic year τ, the value of the IV is defined as:

$$Z_{is\tau} = (\frac{1}{n(J_{s\tau})}) \sum_{j \in J_{s\tau}} [(\frac{1}{\sum_{k \neq s} \sum_{t \neq \tau} \sum_{m} \sum_{r} d_{jkt} n_{ktmr}}) \sum_{k \neq s} \sum_{t \neq \tau} d_{jkt} (\sum_{m} \sum_{r} n_{ktmr} (\bar{P}_{ktmr} - \bar{P}_{mr}))] \tag{9}$$

where j denotes jth principal (or assistant principal); $J_{s\tau}$ is the set of principals in school s in academic year τ;[49] $n(J_{s\tau})$ is the number of principals in school s in academic year τ; d_{jkt} is one if principal j has worked as a principal in school k in academic year t, and zero otherwise; n_{ktmr} is the total number of m type offenses among students' rth offenses in the school k in academic year t, where rth refers to first, second, or third offenses. The summation $\sum_{k \neq s} \sum_{t \neq \tau} \sum_{m} \sum_{r} d_{jkt} n_{ktmr}$ represents the total number of offenses in other schools in which principal j has worked. Let $\bar{P}_{ktmr}$ be the out-of-school suspension rate for the m type misbehavior among students' rth offenses in school k in academic year t, and $\bar{P}_{mr}$ be the total out-of-school suspension rate for the m type misbehavior among students' rth offenses in the sample. Thus, $\bar{P}_{ktmr} - \bar{P}_{mr}$ normalizes the punishments by the most important features (r and m) of offenses and captures the relative harshness of punishments. With this normalization the IV is less likely to reflect the types of offenses that the principals faced in other schools and, thus, more likely to reflect their preferences for or beliefs about the

[49]For most of the academic years, I do not observe the exact dates that a principal works in a school. I define principal teams by academic year, which means that principals defined to be on one team could have worked in non-overlapping periods of the academic year. I use at most two principals and at most five assistant principals per school/year. If, in the academic year, there were more principals in the school, I use those who have worked longest time in the school in the academic year (I observe an approximate length of time that each principal works in a school in an academic year).

harshness of punishments.[50]

As reported in Table 1, there are some categories (or types) of offenses that were detected or reported in only a small number of schools. I find that these categories (or types) of offenses weaken the performance of the IV. To achieve better performance, I construct the IV using that part of the data for which the observed punishment is applied to well detected and reported offenses, namely the "violence" and "disrespect" category of offenses.[51] The IV is missing in some schools in some academic years (12 percent of all student-year observations) since there were no principals or assistant principals who had transfer experience. These student-year observations are not used for the corresponding estimation.

In addition to the IV defined above (henceforth called the main IV), I find that instrumental variables constructed for specific categories (or types) of offenses may have better first stage performance for these categories (or types) of offenses. Therefore, I use these instrumental variables for corresponding specifications.[52] For estimating equation (2), I use two instrumental variables consisting of the out-of-school suspension rates for only the first offense ($r = 1$) and only the second offense ($r = 2$), respectively.

Conditional on the control variables, the student unobserved heterogeneity and the school fixed effects, estimation requires that IV affects the student's behaviors or achievement outcomes only through the principal teams' punishment decisions in the current school. The sorting of principals into schools may be a concern. For example, tough principals may be more (or less) likely to be selected into schools with worse qualities or with more disruptive students. Since I control for time-varying unobserved heterogeneity of students, school fixed

[50]The IV is analogous to the average of all principal team members' average regression residuals, where the out-of-school suspension indicator for each offense (in other schools) is regressed on the $r \times m$ offense indicators. An alternative approach constructing the IV is to run a regression (normalization) that also includes all other explanatory variables and the school fixed effects to eliminate their effects on the IV. I do not take this approach in order to avoid the "noise" introduced by the functional form of the regression, which weakens the effectiveness of the IV. Therefore, the IV may also reflect the information found in these explanatory variable, rather than the principals' "beliefs" or "preferences." Because of this, I include all of these explanatory variables and school fixed effects in the estimation to validate the IV.

[51]The IV is constructed by using data from 2008-2009 academic year to 2014-2015 academic year. I only use offenses with $r <= 3$ to eliminate outliers.

[52]Using multiple IVs in these situations is typically less effective than using the best one.

effects and time-varying observed characteristics of schools, the concern stems from time-varying unobserved school factors that affect both students' misbehavior decisions and the assignment of the principal teams in the current academic year.

Because the concern cannot be directly tested, I explore indirect evidence, such as the correlation between the IV and the time-varying observed characteristics of the school, to determine the magnitude of the problem. I regress the main IV on time-varying observed characteristics of the school and school fixed effects (Table 3). The coefficient column shows that, there are only three regressors that are statistically significant at the 5 percent significance level. These are "other disciplinary infraction cases," "total number of classroom teachers," and "PBIS Exemplar school."[53] The magnitudes of the coefficients are small. The F-statistics of jointly testing the significance of all the time-varying observed characteristics is 1.55 with p-value 0.0069. The p-value increases to 0.0974 if the test does not include "other disciplinary infraction cases," and "total number of classroom teachers" variables. The correlation between school time-varying observables and the IV suggests that there is no strong evidence that time-varying unobserved characteristics are correlated with the IV. Since the regressors include many school quality measures in the academic year of concern, the result also suggests that it is less likely that an administrator's out-of-school suspension propensity reflects her other abilities that could directly affect students' behavior or achievement outcomes. However, the fact that there are some significant correlations indicates that including the time-varying observed school characteristics is important for the validity of the IV. The web appendix shows that, for other constructed IVs, used in separate estimations for different categories (or types) of misbehaviors, the correlation between the IV and the time-varying observed school characteristics is generally smaller.

I expect the effect of the IV on the out-of-school suspension decision to be positive, since "tougher" principal teams (proxied by their out-of-school suspension rate in other schools) should be more likely to use out-of-school suspension in the current school. As discussed

[53]PBIS means positive behavior intervention and support.

in the last section, the monotonicity assumption of the IV suggests that the effect (of each margin) of the IV on the punishment for each offense should be non-negative. That is, the "tougher" principal team would be more likely (or equally likely) to use out-of-school suspension for any offenses. While the assumption cannot be directly tested, indirect evidence, such as the effects of the IV on the out-of-school suspension decisions for different types of offenses, could be used to infer the plausibility of the assumption. To check the monotonicity assumption and the first stage performance of the IV, I run the following OLS regressions:

$$P_{is\tau} = \gamma_0 + \gamma_1 X_{is\tau} + \gamma_2 Z_{is\tau} + \phi_s^{sch} + \theta_{\hat{g}_i s\tau} + \epsilon_{is\tau}^z \tag{10}$$

where $P_{is\tau}$ is the punishment assigned for the first offense of student i in school s in academic year τ, and $Z_{is\tau}$ is the main IV;[54] Recall that $X_{is\tau}$ is the vector of control variables; ϕ_s^{sch} is the school fixed effects; $\theta_{\hat{g}_i s\tau}$ is the student's unobserved heterogeneity; and $\epsilon_{is\tau}^z$ is the error term. To infer the monotonicity of the IV, I separately run the first stage regression for each type of offenses to check the sign of γ_2, which I expect to be positive. Table 4 shows the coefficients (γ_2) for each of the regressions with the type of offense listed in the first column. I only include the types of offenses with more than 4000 observations, since all of the coefficients for the types of offenses with less than 4000 observations are not statistically significant. The coefficient column shows that the IV performs well for most type of offenses, especially for the types of offenses in the "violence" and "disrespect" categories. At the 5 percent significance level, the IV has positive effects for twelve types of offenses. There are two types of offenses with negative effects at the five percent significance level. These are "excessive tardiness" and "late to class." As discussed in the data section, it is very likely that consequences for these categories were not well reported. The reporting issue might be the reason that the IV has negative effects on the rate of out-of-school suspension in these categories. Therefore, to improve the plausibility of the monotonicity assumption, I do not include students with these

[54] I only use the punishment for first offenses because a large part of my first stage regressions only involve the punishment for the first offense. The web appendix shows that the results are generally consistent by using the first, the second, and the third offenses of students in the academic year.

two types of offenses in the offender group in the related specifications. I separately estimate the deterrence effect for the "excessive tardiness" offense by using the IV constructed using suspension decisions for "excessive tardiness" offenses only, which has a positive first stage coefficient.[55] In the web appendix, I provide evidence that instrumental variables that are constructed using the same category offenses only have better performance in terms of the monotonicity for these categories. This finding is one motivation for separately estimating the deterrence effects by offense categories.[56]

The last row of the table shows the coefficient, γ_2, from the regression that uses observations of all types of misbehaviors, which informs the first stage regression for some of my specifications.[57] The F-statistic for testing $\gamma_2 = 0$ in the regression is 86, which is well above the rule of thumb for testing weak instruments.[58]

4.2 Disciplinary Punishment Index Construction

In equations (1), (2) and (4), potential punishments for misbehaviors ($P^*_{1is\tau}$, $P^*_{2is\tau}$, $P^*_{is\tau}$) are not observed for each student. As mentioned in section 3.2, I use "normalized" rates of out-of-school suspension within a school/year ($\hat{P}_{1s\tau}$, $\hat{P}_{2s\tau}$ or $\hat{P}_{is\tau}$) to approximate them. I refer to these proxies as "Disciplinary Punishment Indexes (DPI)." The DPI for the wth (1st or 2nd) offense of a student in school s in academic year τ is defined by:

$$\hat{P}_{ws\tau} = \frac{1}{\sum_m n_{wms\tau}} [\sum_m n_{wms\tau} (\bar{P}_{wms\tau} - \bar{P}_{wm})] \tag{11}$$

where $n_{wms\tau}$ denotes the number of type m offenses among students' wth offenses in school s in academic year τ; $\sum_m n_{wms\tau}$ calculates the total number of wth offenses in school s

[55] I did not find a valid IV for the "late to class" offenses.

[56] I also do robustness checks for the effects of suspension on achievement outcomes by separately estimating the effects using different categories of misbehaviors. The results are in the web appendix.

[57] The estimation is with "type of misbehavior" as an additional control variable. The estimation results only change slightly between including or not including "excessive tardiness" and "late to class" offenses.

[58] Since first stage estimations are different across my different empirical models, I report these first stage results separately with my other estimation results in the following subsections or in the web appendix.

in academic year τ; $\bar{P}_{wms\tau}$ is the out-of-school suspension rate for m type offenses among students' wth offenses in school s in academic year τ; and $\bar{P}_{wm}$ is the out-of-school suspension rate for m type offenses among wth offenses in the sample. Similar to the construction of the IV, the normalization allows the measure to reflect the severity of the punishments in the school but not the severity of offenses in the school.

$\hat{P}_{is\tau}$ is constructed in an analogous manner by using the out-of-school suspension decisions for all the first, second and third offenses of students in an academic year.[59] The DPI for each category of offenses and some selected types of offenses are also constructed by using the out-of-school suspension decisions for same category (or type) offenses of students. In addition, I also separately construct DPI for black and white students by only using the punishments for black and white students, respectively.[60]

The DPI is used as a proxy variable for the potential punishment for each student in the school in the academic year. For the schools without offenses, the DPI is missing. This unobservability suggests that my estimation results for general deterrence effects might not be applicable to the schools with no offense cases in the academic year. The DPI may not be a good proxy if it is constructed by only few observations. Therefore, in most of the estimation specifications, I drop schools with less than five offenses.

4.3 Student Unobserved Heterogeneity

As mentioned in section 3.2, I control for student unobserved heterogeneity to address the selection of students into the offending group because of punishment differences. I use a two-step approach to capture the student unobserved heterogeneity. The first step is to partition all I students into G groups. In the second step, the group-specific unobserved heterogeneity is estimated separately in equations (1) though (4) with all other parts of the

[59] As discussed in the last section, the offenses with type "excessive tardiness" and "late to class" are not included in this calculation.

[60] The DPI for all offenses or the DPI at the category level typically have a mean close to zero, and standard deviations from 0.2 to 0.4. Note that the range of DPI could be larger than 1, but the DPI values that are outside a range of 1 are rare cases.

models by imputing the estimated group memberships of students into the models.

Specifically, in the first step, the partition problem is:

$$\min_{g(1),\ldots,g(I),\theta_1,\ldots,\theta_G} \sum_{i=1}^{I}\sum_{\tau=1}^{\bar{\tau}}\sum_{m=1}^{M}(Y_{i\tau}^{m}-\hat{\beta}_{\tau}^{m}X_{i\tau}-\hat{\phi}_{s\tau}^{sch,m}-\theta_{g_i\tau}^{m})^2 \tag{12}$$

where $g(i)$ indicates that student i is assigned to group g; later, I use g_i to denote the assigned group of student i. The vector $\theta_g = (\theta_{g1}^1,\ldots,\theta_{g1}^M,\ldots,\theta_{g\bar{\tau}}^1,\ldots,\theta_{g\bar{\tau}}^M)$ contains mean value of each feature in each academic year among students in group g. The mean value of the mth feature in academic year τ among students in group g is $\theta_{g_i\tau}^m$.

The mth feature of student i in academic year τ is defined by the expression $Y_{i\tau}^m - \hat{\beta}_{\tau}^m X_{i\tau} - \hat{\phi}_{s\tau}^{sch,m}$. It requires that I regress the mth outcome of student i in academic year τ, $Y_{i\tau}^m$, on a vector of control variables, $X_{i\tau}$, and school-year-category fixed effects in order to obtain estimated values, $\hat{\beta}_{\tau}^m$ and $\hat{\phi}_{s\tau}^{sch,m}$.[61] The mth outcome, $Y_{i\tau}^m$, is the frequency of the mth category offense of student i in academic year τ. The regressions are run separately for each category of offenses, m, in each academic year, τ. The expression $Y_{i\tau}^m - \hat{\beta}_{\tau}^m X_{i\tau} - \hat{\phi}_{s\tau}^{sch,m}$ equals the regression residuals, which are used to partition students.

The regressions are similar to the reduced form of equations (1) and (2), but the dependent variables (the frequencies of each category of offenses of a student within an academic year) aggregate the information of the "first offense" and "second offense" and gains further efficiency by using information on third and more offenses of this student in the academic year.[62] The regression residuals reflect the additional factors that influence a student's misbehaviors (i.e., factors other than observed characteristics of the student and observed and unobserved characteristics of the school). Common patterns among these additional factors are used to identify a student's group membership. That is, they reflect students' unobserved types, such as sensitivity to punishments.

[61] If student i has transfer experience within academic year τ, her school, s, is defined by the one in which most of her offense records come from.

[62] Note that the school-year-category fixed effect soaks up all information in the IV in the reduced form estimation of equations (1) and (2).

The partition problem is solved by a k-means algorithm that proceeds by alternating between "assignment" and "updating." The "assignment" step assigns group memberships for all students; the "updating" step updates the new mean, θ, of each group. The group membership of student i is solved by:

$$\hat{g}_i = \underset{g \in \{1,\ldots,G\}}{\operatorname{argmin}} \sum_{\tau=1}^{\bar{\tau}} \sum_{m=1}^{M} (Y_{i\tau}^m - \hat{\beta}_\tau^m X_{i\tau} - \hat{\phi}_{s\tau}^{sch,m} - \theta_{g\tau}^m)^2 \tag{13}$$

The partition task is designed to classify students by $M \times \bar{\tau}$ features captured by the regression residuals. The fact that data are not available for all students in all academic years creates a complication. Only using students with all years of data may create a selection problem if, for example, students who were offenders or who were suspended, have fewer years of data. Therefore, I group the students by the number of academic years they contribute, and separately classify students within the same "observed years of data" group into more groups defined by the algorithm. The classification may be not fully efficient because it forces students with different numbers of academic years of data to be classified into different groups. However, following the spirit of Bonhomme and Manresa (2015), the fact that it generates more groups reduces the biases of the common parameters that I estimate.[63]

I classify each sample of students (with the same observed years of data) into ten groups. Since there are five years of data in the sample, I have fifty groups of students. Because the k-means algorithm is sensitive to the choice of initial value, I use 500 random initial values to select the classification with the best performance. About 45 percent of students contribute data all five academic years; I use this sample to illustrate the classification results. In total,

[63]My sample of students are followed for five academic years. In a robustness check, I find that a partition task that only uses students with three or more academic years of data produces similar final estimation results to the discussed task. Bonhomme et al. (2016a) suggest that, based on a set of assumptions, different moments and specifications could be used as features for the partition when the moments provide information about the student's group membership. Therefore, I could also add, for example, academic achievement outcomes of students as additional features. Bonhomme et al. (2016b) also show that, with bias reduction, the clustering method could achieve satisfying identification without finite group assumption of the heterogeneity.

there are twelve categories of offenses used for the classification; therefore, for students with five academic year data, their classification is based on sixty ($M \times \bar{\tau}$) features.

In Figure 1, I show trends of the frequencies of each category offenses across years for each group of students. Each graph represents each of the twelve categories of offenses. Each line color represents each of the ten student groups. Students in different groups are observed to behave quite differently, which shows the efficiency of the classification. The group represented by the yellow line has low numbers for each category of offenses, suggesting that it is a rarely misbehaved group of students; in contrast, some other groups of students misbehave frequently across each category of offenses or across years. These groups might capture the groups of students that are not sensitive to punishments. There are groups of students with offense frequencies that fluctuate intensively across academic years (or across offense categories); therefore, some of these groups might capture the "punishment compliers."

5 Results

In this section, I present estimation results for equations (1) through (4). In addition to the results from my preferred estimation strategy, I also show results from other specifications that incrementally address the endogeneity and selection issues so that we can learn about the source of any bias.

5.1 Results for Deterrence Effects

Table 5 reports estimates of α_{11} in equation (1) – the "general deterrence effects" for the first offense.[64] While the rows show the results for different categories of misbehavior, the columns compare the results from different estimation strategies. I begin with results for "all

[64]The estimates for other coefficients of equation (1) are reported in the web appendix.

offenses" (row 1), which should be interpreted with caution for the following reasons.[65] First, the estimates may reflect the effects for some types of misbehaviors that mostly contribute to the variations of the DPI.[66] Second, the estimates may reflect the effects for offense types punished by discipline decisions that are mostly affected by the IV.[67] In rows 2-6, I report the results for different categories of misbehavior.[68]

The first method (OLS) is an OLS model with all control variables and school fixed effects. The second method (OLS&GFE) adds controls for student unobserved heterogeneity; I use the term "grouped fixed effects" (GFE) to represent the student unobserved heterogeneity, which follows the terminology used by Bonhomme and Manresa (2015). The estimates from these two methods suggest that the "general deterrence effect" is statistically significant for most of the categories. The OLS&GFE estimates are slightly, but statistically significantly, different from the OLS estimates for "all offenses," "violence" and "disrespect," which shows that the additional controls for student unobserved heterogeneity may address some biases in estimation. However, as discussed before, the "reverse causality" problem and some types of unobserved factors might not be addressed by these controls. Therefore, I further address these issues by instrumenting for the DPI variable, the key explanatory variable. The third method (2SLS) instruments for the DPI variable, but does not include controls for student unobserved heterogeneity; the fourth method (2SLS&GFE) adds controls for student unobserved heterogeneity. The first stage (adjusted) F-statistics for these estimates are all well

[65]For "all offense" sample, the key explanatory variable, the DPI, is constructed by the suspension decisions for students' first offenses in an academic year; the offense could be any type of misbehavior. The dependent variable is the indicator that the student committed first offense (any type) in the academic year. As discussed in section 4.1, the offense types "excessive tardiness" and "late to class" are not used for constructing the DPI or the offense indicator.

[66]Since the DPI reflects weighted deviations of the schools' suspension rates (for a type of misbehavior) from the average suspension rates in all schools, some serious offenses that all schools assign suspensions to, for example, would contribute little to the variation. In addition, the DPI (for all offenses) for different schools is calculated by different types of misbehaviors due to the different reporting rates, catching rates and offense rates for them across schools, which means that the estimates should be interpreted with caution.

[67]Using "all offenses" is also less likely to satisfy the monotonicity assumption. Although my endogenous variable is continuous, the underlying intuition of monotonicity generalizes to this case.

[68]The DPI measure for a category of misbehavior is constructed by using only the suspension decisions for students' first offenses of the category of misbehavior. The dependent variables are whether or not the student committed the first offense of the category of misbehavior in the academic year.

above the rule of the thumb for testing for weak instruments, and are reported in the web appendix (Stock and Yogo 2005). Note that 2SLS and 2SLS&GFE estimates all suggest higher "general deterrence effects," which are statistically different from the OLS and OLS&GFE estimates. Since harsher discipline rules are more likely to be used for the school/year with higher offense rates or more severe offenses, intuitively, OLS and OLS&GFE estimates might understate the improvement achieved by harsher discipline rules. Therefore, the results suggest that 2SLS and 2SLS&GFE estimation methods provide additional reductions in bias. Another possible explanation is that these IV estimates may capture some "local" effects for some types of offenses or some subpopulations of students if the IV only affects the discipline decisions for them. The 2SLS&GFE estimates are different from 2SLS estimates, which shows that the estimator further addresses some estimation issues.[69]

The last column shows means of the dependent variables, which capture average (reported) offense rates in the estimation sample. By dividing the coefficients by the means, my preferred estimates (from the model 2SLS & GFE) suggest that a 10 percentage point increase in the out-of-school suspension likelihood index could reduce the mean rate of a student committing any type of offense in a year by about 15.6 percent, "violence" offenses by about 13 percent, "disrespect" offenses by about 11.5 percent, "truancy" offenses by about 22 percent, and "drug" offenses by about 18 percent.[70] The estimate is not statistically significant for the "property" category. A possible explanation might be that the low offense rate of this category makes the DPI measure less precise and identification of the effect more difficult (although the empirical framework may address part of the measurement error problem).

[69]The differences are not statistically significant due to the large standard errors of the IV estimates.

[70]Although I use the average reported offense rate to calculate these percentages, interpretation of them may generalize to percentages of real offense rate reduction (observed and unobserved) under the assumption that my empirical strategy effectively addresses the related econometric issues. As discussed in the data section, the results for the drug category should be interpreted with caution because there might be additional punishments for these types of offenses. A robustness check shows that the estimate is not statistically significant for the drug category after controlling for the same grade peers' drug offense rate, which captures the direct general deterrence effects (without effects that work through peer effects). The robustness check results for other categories are also shown in the web appendix.

To further explore the sources of the effects, I estimate α_{11} for different types of misbehaviors, in which the DPI for a type of misbehavior is constructed by using the suspension decisions for the specific offense type only. In Table 6, I show the results for types with the highest offense rates because the estimate is typically not statistically significant for types with lower offense rates. The estimates indicate significant "general deterrence effects" for most of the offense types. The changing patterns of estimates across estimation methods are consistent with the patterns in Table 5. My preferred estimation method (2SLS&GFE) shows that the "general deterrence effects" are heterogeneous across different types of misbehaviors. They are high for offense type "aggressive behavior," "insubordination," and "inappropriate language" – a 10 percentage point increase in the out-of-school suspension likelihood could reduce the mean offense rate for these types by about 30-40 percent. The percent reduction is about 12 percent for "skipping class," 20 percent for "disruptive behavior," 7.5 percent for "disrespect to faculty," and 16 percent for "excessive tardiness."[71] It is not statistically significant for the "fighting" offense. The heterogeneity in response may stem from the differences in motivation, nature, or characteristics of offenders for these offenses. Another possible explanation is that the heterogeneity may be partly due to differences in average out-of-school suspension rates assigned for these offenses, which are shown in Table 1.[72] Finally, these results should be interpreted with caution as the offense rate for each offense type is relatively low, which makes the DPI measure less precise and the identification of the effect more difficult.

To further explore the potential heterogeneity of the general deterrence effect, I present estimates stratified by student observables (Table 7). The DPI measures are constructed by using the suspension decisions for the students with the corresponding observables. I only present the results for the well reported "violence" and "disrespect" categories because DPI

[71] The percent reduction is calculated by dividing the coefficients by the means of dependent variables.

[72] A 10 percentage point increase in the suspension rate might mean a much harsher discipline when the current average suspension rate is low; in addition, the reported average suspension rate may also represent the "severity" of these types of offenses, which may suggest that the heterogeneity partly contributes to the severity of offenses. However, the correlation between the average suspension rate and the general deterrence effect is not quite clear.

measures are less precise and the estimates are more likely to suffer from a weak instruments problem after the sample is conditioned on a particular student characteristic.

The differences between OLS&GFE and 2SLS&GFE estimates are consistent with the patterns discussed in Table 5 and Table 6. My preferred estimates (2SLS&GFE) find statistically significant "general deterrence effects" for most of the student subpopulations, and the effects are heterogeneous. First, I find that "general deterrence effects" are not statistically significant (and with small coefficients) for high school violent behavior, and elementary school disrespectful behavior.[73] Second, I find that the "general deterrence effects" might be higher (in terms of the percent reduction of mean rate) for white students than black students – a 10 percentage point increase in the suspension likelihood index reduces the mean offense rate of "violence" and "disrespect" behaviors for white students by about 16 percent and 24 percent, and by 11 percent and 9 percent for black students. In addition, I find that, in terms of the percent reduction of mean rate, the effects are smaller for female students than male students, smaller for economically disadvantaged students than those not economically disadvantaged, smaller (and not statistically significant) for "violence" offenses of students with lagged math scores below (or equal) the average than those with lagged math scores above the average.[74]

Table 8 reports estimates of the general deterrence effect (α_{21}) and the specific deterrence effect (α_{22}) for students' second offenses in an academic year (equation 2), which are estimated separately by offense category.[75] The table includes results for the better reported "violence" and "disrespect" categories, and for the "truancy" category. I show the results for other categories or types of offenses in Table A1 in the appendix. Most of the OLS and

[73] A robustness check by controlling for offense rates of all types of peers' misbehaviors also finds that the effects are relative low (but statistically significant) for the disrespectful behavior of high school students.

[74] The percent reduction is calculated by dividing the coefficients by the means of dependent variables (the fourth number in each cell). Again, the heterogeneity should be interpreted with caution because of the following reasons. First, some of the differences are not statistically significant. Second, some of them are mostly due to the differences in the mean offense rates. Third, the DPI variables may less precisely reflect the punishment severity in the school because they are constructed with less observations.

[75] The DPI variable, $\hat{P}_{2s\tau}$, is constructed by only using students' second offenses of the category in the academic year.

OLS&GFE estimates suggest statistically significant positive "general deterrence effects" for students' second offenses. However, the estimates for α_{22} suggest that the "specific deterrence effect" is negative for "violence" offenses and it is positive but small for "disrespect" and "truancy" offenses. The OLS&GFE estimate suggests that for "disrespect," for example, the suspension experience itself could only reduce the mean rate of the student's re-offending by less than 2 percent. The 2SLS and 2SLS&GFE estimates suggest that both "general deterrence effects" and "specific deterrence effects" are not statistically significant.[76] However, one problem with these estimates is that the collinearity between α_{21} and α_{22} is severe because the IV method reduces the variation of $\hat{P}_{2s\tau}$ and $P_{1is\tau}$ in the second stage estimation. This reduced variation inflates the standard errors and the magnitude of these estimates; therefore, the inference is less precise and the results should be interpreted with caution.[77]

I further explore the efficacy of "general deterrence," α_{21}, by estimating equation (2) separately for student-observations that were out-of-school suspended for their first offense ($p_{1is\tau} = 1$) and those that were not out-of-school suspended for their first offenses ($p_{1is\tau} = 0$).[78] Table 9 shows the estimates of α_{21} using these specifications for the "violence," "disrespect," and "truancy" categories. I show the results for other categories and types of offenses in Table A2 of the appendix.

My preferred estimates (2SLS&GFE) suggest that the "general deterrence effects" are not statistically significant for students who were out-of-school suspended for their first offenses in the academic year. They are also not statistically significant for the "violence" and "truancy" category offenses for students who were not suspended.[79] In addition, I find that the effect is statistically significant for the second "disrespect" offenses of students who were not suspended for their first offenses. However, the estimate suggests a smaller effect compared to the effect for students' first "disrespect" offenses – a 10 percentage point increase in the

[76] F-tests also show that they are jointly not statistically significant.

[77] Robustness checks, by instrumenting only one of these endogenous variables, yield similar results.

[78] By this specification, I could estimate the "general deterrence effects" without adding $P_{1is\tau}$ as an explanatory variable in the regression, which solves the collinearity problem.

[79] The large coefficients and standard errors for the "truancy" category for these estimates are because the first stage performances of the IV are relative weak, with (adjusted) F-statistics of about 7 and 10.

suspension likelihood index reduces the mean rate of re-offending of "disrespect" offenses by about 8 percent.

5.2 Results for the Effects on Achievement Outcomes

Table 10 reports estimates of α_{31} in equation (3) – the effects of out-of-school suspension experience on offending students' achievement outcomes. The outcomes (dependent variables) include end-of-grade math score, a dropout indicator (finally graduate or drop out from high school) and ACT composite score.[80] The "dropout" dependent variable is whether or not the grade 9 student finally drops out from high school, which may happen in any future grades.[81] I estimate how it is affected by the student's suspension experience in grade 9. In the web appendix, I show that the results are robust to the specifications that use students' suspension experience in grade 9-10 or grade 9-12. I mainly focus on interpreting the results for grade 9 because adding student observations in higher grades may create additional dynamic selection problems, which may make interpretation of the results less clear.[82] For estimation using ACT composite scores as the outcome, I use student-year observations in grade 9-10 who have their final ACT composite scores.[83] In the web appendix, I show that the results are similar when I use student observations in grade 9 only.

Columns 2-6 report results using different estimators where the key explanatory variable is an indicator of whether or not the student was ever suspended out-of-school during the academic year. To compare my methods/results with those from the exiting literature, I begin with OLS regressions that include controls for school-level observables and school

[80]The end-of-grade math scores are normalized to have a zero mean and standard deviation of one among students who took the same tests across the state. The end-of-grade reading scores are used as a robustness check and the results are reported in the web appendix.

[81]In North Carolina, students can drop out of high school at 16 years old.

[82]The problem is that the offending students in grade 10 who did not drop out at the end of grade 9 may be a selected group, such as a relatively better behaved group in the offending student population. Although by assumption, my preferred estimation strategy may address the problem for estimating α_{31} for observed offending students in grade 10 (or higher grade), but the concern that they may be the relatively better behaved group, for example, suggests that extrapolating these estimates to all offending students should be done with caution.

[83]The choice is mainly based on the data availability, which was discussed in section 2.

fixed effects only. The results in Column 2 are from a regression using all students (not only offending students); it compares the achievement outcomes of students who had experienced suspension to a control group that includes both students who had an offense record but were not suspended and students who did not have an offense record in the academic year. A suspension experience decreases grade 3-8 students' end-of-grade math scores by about 0.2 standard deviations. Suspension experience in grade 9 increases the dropout probability by about 18.5 percentage points. Suspension experience in grade 9-10 also lowers ACT composite scores by about 1.1 points. One problem with interpretation of these effects is that, among students in the control group, those who did not commit an offense are not at risk of suspension. To measure the causal effect of suspension on outcomes, we may desire to focus on students facing the possibility of being suspended. Beginning with Column 3, I limit the estimation sample to students who had an offense record in the academic year and label them "offending students."[84] Column 3 reports the results with the same controls as column 2 but uses the "offending students" sample. The results indicate smaller negative effects of suspension for all three achievement outcome measures, which suggests that using a more representative sample may solve some selection issues. To further reduce the endogeneity and selection biases, I add all student-level controls discussed in section 4 and report the results in the "OLS with full controls" column. The results indicate much smaller negative effects of suspension – a suspension experience reduces students' end-of-grade math scores by about 0.013 standard deviations, and ACT composite scores by about 0.21 points. It increases the dropout probability by about 7.2 percentage points. These effects fall even more (to 0.012 standard deviations, 0.17 points, and 5 percentage points) after adding an additional control for student unobserved heterogeneity ("OLS & GFE" column).

When I address the estimation biases by instrumenting for the suspension decision (2SLS and 2SLS&GFE), I find that the effects for all three outcomes are not statistically signif-

[84] Although the sample of offending students does not include students who committed an offense but did not get a record, it should be more representative of the "at risk for punishment" students than the "all students" sample.

icant.[85] The estimates for end-of-grade math and dropout probability also change signs. The results should be interpreted as the effects of suspension on achievement outcomes of students on the margin of suspension - the compliers. The large standard errors of the estimates also suggest using caution in interpretation. However, the results from all estimates consistently show that the negative correlation between suspension experience and students' end-of-grade math scores or dropout probability become smaller once I further address the endogeneity and selection biases. These findings suggest that the correlation is largely (or completely) explained by the correlation between observed or unobserved factors of students (or offenses) and these achievement outcome measures.

Some additional information suggest that the results for ACT scores should be undertaken with caution. First, since the ACT is required for students in grade 11, offending students who drop out before the test do not generate observed test scores. Therefore, the students who have the scores may be a selected group and extrapolating the estimates to all offending students should be done cautiously. An additional concern is that the suspension decision may induce students with different unobserved factors to drop out at different rates, which may imply that the offending students who have both ACT scores and suspension experience have different unobserved factors than offenders who had ACT scores but did not experience suspension.[86] Because the student unobserved heterogeneity is mostly identified by students' misbehavior information but not dropout decisions (which cannot be repeatedly observed), the unobserved heterogeneity in this dimension might not be captured. However, my estimation for the dropout may indicate that this concern is not severe because the suspension experience might not significantly affect students' dropout probability.

In the last two columns of Table 10, I provide the robustness checks with the key explanatory variable being an indicator of whether the student was suspended for her *first*

[85]The first stage (adjusted) F-statistics for the IV are 65 and 74 for the 2SLS and 2SLS&GFE estimates of end-of-grade math test scores, 26 for the 2SLS and 2SLS&GFE estimates of dropout indicator, and 17 for the 2SLS and 2SLS&GFE estimates of ACT composite scores.

[86]In this case, the difference of ACT scores between those who have or did not have suspension experience may be a reflection of these unobserved factors but not the "causal effects" of suspension.

offense in that academic year.[87] As introduced in section 3, the specification controls for students' offense frequencies in the previous academic year instead of in the academic year of concern. These estimates may capture the indirect effect of suspension on achievement outcomes resulting from the offending students' behavior changes in the current academic year. Estimates from the preferred model (2SLS& GFE) suggest that the effects are not statistically significant, which is consistent with the previous findings.[88]

One interesting question is whether the effects of suspension on achievement are different for black and white students. Table 11 provides results from models estimated separately for black and white students. Each of the four models (Columns 2-5) uses the full set of control variables and the indicator of whether the student was ever suspended out-of-school during the academic year as the key explanatory variable. The OLS&GFE estimates suggest that the negative effects of suspension on all three achievement outcomes are larger for white students than for black students. The 2SLS and 2SLS&GFE estimates suggest that the effects of suspension experience on end-of-grade math scores and dropout probability are not statistically significant for either black or white students.[89] The 2SLS estimate suggests that the negative effect on white students' ACT scores is about 3.3 points, which is statistically significant at the 10% confidence level. The 2SLS&GFE point estimate shows that the negative effect on white students' ACT scores is about 2.9 points, but it is not statistically significant.[90]

[87]In the web appendix, I show the results are robust to only using students who were suspended for their first "violence" and "disrespect" offenses.

[88]The large standard errors and point estimates for ACT scores are likely due to a weak instruments problem. The (adjusted) F-statistics of the IV in the first stage regression is only about 3.

[89]The first stage (adjusted) F-statistic for the IV is 50 for black students' end-of-grade math scores, and 7 for white students' end-of-grade math scores. In addition, the first stage (adjusted) F-statistic for the IV is 6 for white students' dropout probability and 13 for black students' dropout probability. The IV is relatively weak for white students; however, in the web appendix, I show that the results are robust by using different specifications for white students.

[90]A robustness check using student unobserved heterogeneity that does not change across time suggest the estimate for white students' ACT scores is statistically significant at 10% confidence level. The same robustness checks are done for all other estimates, and show that other results are robust. The large standard error for 2SLS and 2SLS&GFE estimates for black students' ACT scores is because of the weak instruments problem; the first stage (adjusted) F-statistic for the IV is only about 1. The F-statistic for white students' ACT scores are about 22;

While these results illustrate the effects of suspension experience on offending students' achievement outcomes, I further explore the effects of exclusionary school discipline on achievement outcomes of students with no offense record and the overall effects on all students' achievement. As discussed in the introduction, these effects may encompass students' and peers' behavior changes resulting from the "general or specific deterrence effects" or from the "incapacitation effect." Table 12 reports estimates of α_{41} in equation (4).[91] Again, I separately report OLS, OLS&GFE, 2SLS, and 2SLS&GFE estimates in different columns.[92] Estimates from my preferred model (2SLS&GFE) show that the effect is not statistically significant for end-of-grade math scores, dropout probabilities and ACT composite scores of "all students." In addition, the effect is not statistically significant for dropout probabilities and ACT composite scores of students with no offense record.[93] However, I find that the effect is positive and statistically significant for the end-of-grade math scores of middle school students with no offense record. A 10 percentage point increase in the suspension likelihood index could increase the end-of-grade math scores of "well behaved" middle school students by about 0.02 standard deviations.

To further explore whether the overall effects are heterogeneous by racial group, Table 13 reports estimation results for white and black students by splitting the sample. The 2SLS&GFE estimates suggest that harsher school discipline has a positive overall effect on middle school white students' end-of-grade math scores. A 10 percentage point increase in the suspension likelihood index increases the end-of-grade math scores of middle school white students by about 0.028 standard deviations. I do not find such evidence for middle school

[91] A student with no offense record means that she does not have any offense record in my sample period. I only include students with no offense record at least for three academic years to make sure that they are "well behaved." For the achievement outcome measures "dropout" and "ACT composite scores," I also drop students with any offense record in 2013-2014 and 2014-2015 academic year.

[92] All of the estimates are with a full set of control variables. The key explanatory variable, $\hat{P}_{s\tau}$, is constructed by using "normalized" suspension rates for students' first, second, and third offenses in the academic year. As discussed before, "excessive tardiness" and "late to class" are not included into this calculation.

[93] The larger negative coefficients for ACT composite scores suggest that one should interpret the results for ACT scores with caution.

black students' end-of-grade math scores.[94]

6 Conclusions

Exclusionary school discipline techniques are widely criticized for their inability to improve students' behavior and for their adverse effects on students' achievement outcomes. However, I find that disciplinary rules exhibiting a higher out-of-school suspension likelihood could significantly deter students from committing first offenses. Estimates from my preferred estimation method also suggest that the adverse effects of out-of-school suspension experience on offending students' end-of-grade test scores and high school dropout probability are not statistically significant. Moreover, contrary to Perry and Morris (2014), I find that disciplinary rules with higher out-of-school suspension likelihood could improve the end-of-grade math scores of "well behaved" middle school students. The results imply that policies that reduce or remove suspension options from schools should carefully consider these benefits of the disciplinary practice. Particularly, policies that focus on reducing out-of-school suspension rates for minority groups may increase offense rates of these minority groups. The results also suggest that, contrary to the existing literature (e.g., Morris and Perry 2016), there is no evidence that the suspension disparity between white and black students creates crucial black-white achievement gap.

However, my estimates also suggest that the disciplinary practice is less effective or ineffective for repeat offending students, especially for students who have had suspension experience. Since repeat offenses account for a large portion of all infractions, the results suggest that it is important to find a more effective approach to deal with them. I also find suggestive evidence that the disciplinary practice might be less effective or ineffective for some types of (first) offenses among certain student subpopulations, such as high school students' violent behavior. Therefore, alternative approaches to deal with these offenses are

[94]The (adjusted) F-statistics for the IV in the first stage regression is about 7 for middle school black students' end-of-grade math scores.

also important.

When interpreting the results, several caveats should be kept in mind. First, when heterogeneous effects are in play, my preferred method may be more likely to capture weighted average effects for "compliers" (the students who were, or would be, punished differently because they were assigned to principal teams with different out-of-school suspension propensities).[95] Nevertheless, positive general deterrence effects are consistently found for different categories (or types) of misbehaviors and for different students' subpopulations. The results also consistently show that the documented negative effects of suspension experience on end-of-grade math scores or dropout probability are largely due to endogeneity or selection.

Second, since students' out-of-school misbehaviors are not observed in my data, the analysis does not capture the potential effects of exclusionary school discipline in that dimension. Therefore, my estimates may understate the costs of suspensions (or overstate the benefits of them) if they lead to increases in these unobserved misbehaviors.[96]

In addition, since I simplify the punishment as "out-of-school suspension" or not and most of the out-of-school suspensions are short-term, the results for the effects of suspension experience on achievement outcomes are more likely to capture the effects from comparing short-term out-of-school suspensions with other less severe punishments. Therefore, the results may not completely capture the achievement loss of students who were suspended for a long time, and they might also not represent the effects of out-of-school suspension compared to no punishment at all.

[95]The effects on these students may be particularly interesting for policy making, as they may be the student group for which different policies suggest different punishments.

[96]One example is that, compared to the alternative punishments (e.g., no punishment, detention, in-school suspension), out-of-school suspension might be more likely to lead students to commit offenses off-campus. Then, my estimation does not capture this part of social costs. If these off-campus offenses substitute the students observed (detected and reported) in-school offenses, my estimates might also overstate the deterrence effects for in-school misbehaviors.

Table 1: Summary Statistics for Selected Categories of Offenses

Offense	Total Number of Incidents	Percent of Offending Students who receive OSS for 1st offense	for 2nd offense	Percent of Schools with any reported offense among academic years
Offense Category				
Violence	584,566	63.6	64.3	93.0
Disrespect	902,248	29.4	34.9	86.4
Truancy	343,896	17.8	24.7	48.7
Drug	75,072	59.1	58.1	42.0
Property	63,075	53.1	58.2	71.0
Other	1,981,559	17.2	17.2	91.2
Offense Type				
Fighting	198,547	84.9	83.4	77.3
Aggressive Behavior	189,695	44.9	47.2	75.8
Disrespect to Faculty	206,862	34.5	40.8	67.5
Insubordination	435,622	27.5	30.5	62.1
Inappropriate Language	259,337	35.5	41.9	77.3
Skipping Class	204,914	14.4	20.9	38.1
Disruptive Behavior	753,064	23.6	25.2	82.3
Excessive Tardiness	309,018	8.0	8.9	24.3

Note: Column 2 lists the total number of incidents for each category (or type) of offense in the sample. Column 3 (Column 4) lists the percentages of students in the academic year who were punished by out-of-school suspension (or expulsion) for the same offense. Column 5 lists percentages of schools with any reported offense in each category among academic year. The total number of school-year observations is 11,425. Type "fighting" and "aggressive behavior" are in the "violence" category; type "disrespect to faculty," "insubordination," and "inappropriate language" are in the "disrespect" category; type "skipping class" is in the "truancy" category; type "disruptive behavior" and "excessive tardiness" are from the "other" category.

Table 2: Sample Means of Student Characteristics

	Student-Year Sample with		
	No Offense Record	Any Offense Record	
		No OSS	Any OSS
Race			
White	0.574	0.503	0.325
Black	0.228	0.326	0.514
Hispanic	0.118	0.107	0.094
Asian	0.030	0.010	0.006
Multi-Racial	0.035	0.038	0.037
American Indian	0.014	0.014	0.023
Other Race	0.001	0.000	0.001
Disability			
No Disability	0.875	0.829	0.778
Physical Disability	0.056	0.078	0.088
Intellectual Disability	0.069	0.093	0.133
Other Dichotomous Characteristics			
(omitted: alternative group)			
Female	0.531	0.390	0.313
Economically Disadvantaged	0.453	0.603	0.735
Limited English Proficiency	0.062	0.049	0.049
Academically and Intellectually Gifted - Reading	0.143	0.074	0.032
Academically and Intellectually Gifted - Math	0.155	0.081	0.037
Old in the Grade	0.118	0.187	0.301
Repeating Grade in the Academic Year	0.016	0.039	0.104
Mean of Lagged Scores			
Lagged Normalized Math Score	0.043	-0.048	-0.202
Lagged Normalized Reading Score	0.064	-0.041	-0.198
Lagged Score Missing Indicator	0.216	0.152	0.178
Grade level			
Grade 3	0.122	0.047	0.036
Grade 4	0.120	0.055	0.048
Grade 5	0.115	0.063	0.060
Grade 6	0.098	0.113	0.110
Grade 7	0.093	0.122	0.127
Grade 8	0.092	0.124	0.135
Grade 9	0.098	0.141	0.188
Grade 10	0.091	0.126	0.129
Grade 11	0.086	0.113	0.097
Grade 12	0.086	0.096	0.070
Observations (student years)	4,034,542	658,611	577,886

Note: This table separately reports summary statistics for the student-year samples without offense records, with offense records but without out-of-school suspension records, and with both offense records and any out-of-school suspension records. For grade 10-12 students, End-of-Course Test English 1 is used for the calculation of lagged reading scores; End-of-Course Test Algebra 1 is used for the calculation of lagged math scores. Lagged test scores are normalized to have a zero mean and standard deviation of one among the students who took the same tests across the state.

Table 3: IV and Time Varying School Observables

Dependent Variable: Principal Team's Exclusive OSS Tendency (IV)	Coefficient	Standard Error
Violent crime cases last year (N)	0.0009	(0.0006)
Students involved in misbehavior last year (N)	−0.0001	(0.0001)
Students assigned OSS or expulsion last year (N)	0.0000	(0.0001)
Assault, robbery or sexual offense cases (N)	−0.0004*	(0.0002)
Threat or possession of a weapon cases (N)	−0.0003	(0.0004)
Disorderly conduct or harassment cases (N)	−0.0001	(0.0001)
Other violent cases (N)	−0.0000	(0.0001)
Drug related cases (N)	0.0000	(0.0002)
Disrespect cases (N)	−0.0000	(0.0000)
Disruptive behavior cases (N)	0.0000	(0.0000)
Truancy cases (N)	0.0000	(0.0000)
Tardiness cases (N)	0.0000	(0.0000)
Property cases (N)	0.0004	(0.0003)
Other rule violation cases (N)	0.0000	(0.0000)
Other disciplinary infraction cases (N)	0.0001**	(0.0000)
Minor (average OSS days < 0.55) misbehavior cases (N)	−0.0000	(0.0000)
Moderate ($0.55 \leq$ average OSS days ≤ 1) misbehavior cases (N)	−0.0000	(0.0000)
Major (average OSS days > 1) misbehavior cases (N)	−0.0001	(0.0000)
Ratio of black students	0.0056	(0.0327)
Ratio of Hispanic students	−0.0435	(0.0455)
Ratio of other minority students	−0.1419*	(0.0755)
School mean of normalized math score last year	−0.0030	(0.0161)
School mean of normalized reading score last year	−0.0001	(0.0141)
Proportion of students – math scores 2 sd below state average last year	−0.0584	(0.1001)
Proportion of students – reading scores 2 sd below state average last year	0.0799	(0.1011)
Title I eligible school	0.0046	(0.0075)
School-wide title I	0.0131*	(0.0075)
Ratio of teachers licensed in the school for more than 5 years	0.0132	(0.0168)
Ratio of female personnels	0.0579	(0.0420)
Ratio of black personnels	0.0474	(0.0479)
Ratio of non-white non-black personnels	0.0507	(0.0866)
Magnet School Indicator	−0.0044	(0.0167)
Total student number	0.0000	(0.0000)
Students who are economically disadvantaged %	0.0069	(0.0284)
Total full-time equivalent classroom teachers	0.0002	(0.0006)
Total number of classroom teachers	−0.0012**	(0.0005)
Fully licensed teachers %	−0.0153	(0.0432)
Teachers with experience 4-10 years %	−0.0063	(0.0296)
Teachers with experience more than 11 years %	−0.0613*	(0.0316)
Teachers with Advanced Degrees %	−0.0220	(0.0314)
Teacher Turnover Rate %	0.0119	(0.0262)
Average daily school attendance %	0.0089	(0.2181)
Students per Instructional Computer (N)	−0.0009	(0.0008)
Books per Student (N)	−0.0001	(0.0002)
Average age of Books in library or media center	−0.0000	(0.0000)
Classes taught by highly qualified teachers %	−0.0531	(0.0475)
Adequate yearly progress target met %	−0.0120	(0.0120)
Classrooms connected to the Internet %	−0.0539*	(0.0288)
PBIS - Green Ribbon School	0.0059	(0.0072)
PBIS - Model School	0.0029	(0.0079)
PBIS - Exemplar School	0.0229**	(0.0114)
One or more school variables were missing	−0.0026	(0.0100)
Number of school-year observations	9210	

Robust standard errors are in the parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: The Monotonicity of the Instrumental Variable

Type of Offense	Coefficient	Standard Error	Observations
Assault on student	0.133***	(0.039)	16153
Assault on student (no weapon, no serious injury)	0.107*	(0.056)	10244
Fighting	0.022***	(0.009)	156787
Aggressive behavior	0.062***	(0.014)	133864
Bullying	0.143***	(0.031)	30541
Gang activity	0.122*	(0.071)	5552
Disorderly conduct	0.057	(0.038)	23297
Communicating threats	0.061*	(0.033)	20902
Harassment - verbal	0.023	(0.044)	14629
Harassment - sexual	0.077	(0.050)	11803
Disrespect of faculty/staff	0.058***	(0.014)	134443
Inappropriate language/disrespect	0.034***	(0.011)	181037
Insubordination	0.072***	(0.010)	234486
Disruptive behavior	0.020***	(0.007)	368301
Possession of marijuana	−0.062	(0.045)	10707
Possession of a weapon (not firearms or explosives)	0.055	(0.039)	13295
Possession of tobacco	−0.008	(0.041)	11969
Use of tobacco	−0.004	(0.030)	19997
Theft	0.022	(0.026)	35716
Property damage	−0.075*	(0.042)	16329
Inappropriate items on school property	0.033	(0.043)	16779
Skipping class	0.019*	(0.010)	131950
Truancy	0.145***	(0.044)	21672
Leaving class without permission	0.044*	(0.023)	33621
Leaving school without permission	0.004	(0.033)	22644
Skipping school	−0.054*	(0.031)	27448
Late to class	−0.075***	(0.012)	67915
Excessive tardiness	−0.119***	(0.009)	128301
Excessive display of affection	0.045	(0.042)	10145
Honor code violation	0.248***	(0.030)	16427
Dress code violation	−0.016	(0.020)	45371
Falsification of information	−0.037	(0.055)	8683
Being in an unauthorized area	−0.012	(0.032)	21765
Cell phone use	−0.027*	(0.015)	71203
Bus misbehavior	−0.003	(0.006)	131022
Other School Defined Offense	0.093***	(0.019)	86391
Other	0.065***	(0.019)	64154
Misuse of school technology	0.019	(0.047)	11770
For all types	0.037***	(0.004)	1077758

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports OLS regression results of coefficients γ_2 in equation (10) for different types of offenses. The dependent variable is the out-of-school suspension indicator. The key explanatory variable is the (main) instrumental variable. Standard errors are reported in parentheses and clustered at the student level.

Table 5: General Deterrence Effects for the First Offense by Offense Category

Offense Category	OLS	OLS &GFE	2SLS	2SLS &GFE	Sample Size	Mean of Dependent Var.
All Offenses	-0.244^{***}	-0.248^{***}	-0.372^{***}	-0.397^{***}	4545364	0.250
	(0.001)	(0.001)	(0.034)	(0.028)		
Violence	-0.041^{***}	-0.039^{***}	-0.141^{***}	-0.116^{***}	4372421	0.083
	(0.001)	(0.001)	(0.027)	(0.022)		
Disrespect	-0.064^{***}	-0.071^{***}	-0.114^{***}	-0.121^{***}	4003540	0.104
	(0.001)	(0.001)	(0.017)	(0.014)		
Truancy	-0.061^{***}	-0.061^{***}	-0.139^{***}	-0.154^{***}	2981490	0.068
	(0.001)	(0.001)	(0.030)	(0.029)		
Drug	-0.001^{**}	-0.001	-0.038^{**}	-0.040^{***}	2321145	0.022
	(0.001)	(0.001)	(0.015)	(0.014)		
Property	-0.001^{**}	-0.000	-0.009	0.020	2870740	0.015
	(0.000)	(0.000)	(0.065)	(0.031)		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{11} in equation (1) for different categories of misbehaviors (in different rows), estimated separately. Columns 2-5 show the results using different estimation methods. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include the control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. The sample does not include schools with less than 5 observations of students' first offenses in an offense category in the academic year. The mean of the dependent variable (offense indicator) is reported in the last column. Standard errors are reported in parentheses and clustered at the student level.

Table 6: General Deterrence Effects for the First Offense by Offense Type

Offense Category	OLS	OLS &GFE	2SLS	2SLS &GFE	Sample Size	Mean of Dependent Var.
Fighting	-0.009^{***}	-0.009^{***}	-0.019	-0.017	4009898	0.038
	(0.001)	(0.001)	(0.032)	(0.028)		
Aggressive Behavior	-0.018^{***}	-0.018^{***}	-0.242^{***}	-0.166^{***}	3493914	0.037
	(0.000)	(0.000)	(0.050)	(0.032)		
Disrespect to Faculty	-0.024^{***}	-0.026^{***}	-0.029^{***}	-0.031^{***}	3239179	0.040
	(0.001)	(0.001)	(0.010)	(0.009)		
Insubordination	-0.051^{***}	-0.053^{***}	-0.232^{***}	-0.228^{***}	3365506	0.068
	(0.001)	(0.001)	(0.038)	(0.033)		
Inappropriate Language	-0.016^{***}	-0.018^{***}	-0.168^{***}	-0.144^{***}	3833696	0.046
	(0.001)	(0.001)	(0.020)	(0.017)		
Skipping Class	-0.037^{***}	-0.037^{***}	-0.054	-0.056^{*}	2807283	0.047
	(0.001)	(0.001)	(0.034)	(0.034)		
Disruptive Behavior	-0.085^{***}	-0.091^{***}	-0.242^{*}	-0.203^{***}	3813508	0.096
	(0.001)	(0.001)	(0.128)	(0.074)		
Excessive Tardiness	-0.071^{***}	-0.072^{***}	-0.180^{***}	-0.124^{**}	1513515	0.073
	(0.001)	(0.001)	(0.059)	(0.061)		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{11} in equation (1) for different types of misbehaviors (in different rows), estimated separately. Columns 2-5 show the results using different estimation methods. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include the control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. The sample does not include schools with less than 3 observations of students' first offenses in an offense type in the academic year. Offense type "fighting" and "aggressive behavior" are from the "violence" category; offense type "disrespect to faculty,"‘insubordination," and "inappropriate language" are from the "disrespect" category; offense type "skipping class" is from the "truancy" category; offense type "disruptive behavior" and "excessive tardiness" are from the "other" category. The mean of the dependent variable (offense indicator) is reported in the last column. Standard errors are reported in parentheses and clustered at the student level.

Table 7: General Deterrence Effects for the First Offense by Student Characteristics

Sample	Violence		Disrespect	
	OLS &GFE	2SLS &GFE	OLS &GFE	2SLS &GFE
Elementary School	-0.027***	-0.108***	-0.030***	-0.024
	(0.001)	(0.014)	(0.001)	(0.023)
	1146718	1146718	854611	854611
	0.069	0.069	0.048	0.048
Middle School	-0.046***	-0.152**	-0.076***	-0.158***
	(0.002)	(0.062)	(0.001)	(0.045)
	1461682	1461682	1433382	1433382
	0.124	0.124	0.111	0.111
High School	-0.013***	-0.004	-0.070***	-0.167***
	(0.002)	(0.047)	(0.001)	(0.019)
	1880429	1880429	1879421	1879421
	0.056	0.056	0.117	0.117
White Students	-0.024***	-0.079***	-0.050***	-0.168***
	(0.001)	(0.025)	(0.001)	(0.021)
	2241355	2241355	2040484	2040484
	0.054	0.054	0.072	0.072
Black Students	-0.054***	-0.175***	-0.101***	-0.176***
	(0.002)	(0.059)	(0.002)	(0.032)
	1229951	1229951	1150650	1150650
	0.149	0.149	0.180	0.180
Female	-0.016***	-0.003	-0.047***	-0.048***
	(0.001)	(0.050)	(0.001)	(0.015)
	1896881	1896881	1744554	1744554
	0.055	0.055	0.077	0.077
Male	-0.050***	-0.173***	-0.083***	-0.174***
	(0.001)	(0.033)	(0.001)	(0.021)
	2279396	2279396	2106486	2106486
	0.113	0.113	0.135	0.135
Econ Disadvantage	-0.042***	-0.164***	-0.084***	-0.145***
	(0.001)	(0.036)	(0.001)	(0.024)
	2231892	2231892	2060519	2060519
	0.121	0.121	0.146	0.146
Not Econ Disadvantage	-0.018***	-0.070**	-0.043***	-0.114***
	(0.001)	(0.029)	(0.001)	(0.017)
	2040170	2040170	1854851	1854851
	0.045	0.045	0.060	0.060
Lagged math ≤ 0	-0.034***	-0.133	-0.086***	-0.194***
	(0.002)	(0.082)	(0.002)	(0.045)
	832128	832128	794255	794255
	0.135	0.135	0.184	0.184
Lagged math > 0	0.028***	-0.079***	-0.054***	-0.078***
	(0.001)	(0.026)	(0.001)	(0.016)
	2654784	2654784	2454586	2454586
	0.073	0.073	0.082	0.082

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{11} in equation (1) for the "violence" and "disrespect" categories by student characteristics, estimated separately. OLS&GFE: a model with all control variables, school fixed effects and student unobserved heterogeneity; 2SLS&GFE: additionally instruments the DPI variable. Lagged math score is normalized to have a zero mean and standard deviation of one among the students who took the same tests across the state. Standard errors are reported in parentheses and clustered at the student level. The third number in each cell is the sample size. The fourth number in each cell is the mean of dependent variable (offense indicator). The sample for each estimation does not include schools with less than 3 students' first offense observations from the category in the subpopulation in the academic year.

Table 8: General and Specific Deterrence Effects for the Second Offense

Offense Category	OLS	OLS &GFE	2SLS	2SLS &GFE	Sample Size	Mean of Dependent Var.
Violence						
α_{21}	-0.028^{***}	-0.034^{***}	-2.437	-2.489	351829	0.280
	(0.005)	(0.005)	(3.594)	(4.660)		
α_{22}	0.008^{***}	0.004^{**}	2.515	2.582		
	(0.002)	(0.002)	(3.901)	(5.081)		
Disrespect						
α_{21}	-0.064^{***}	-0.069^{***}	1.951	2.122	407292	0.419
	(0.005)	(0.005)	(2.498)	(3.028)		
α_{22}	-0.000	-0.008^{***}	-1.780	-1.915		
	(0.002)	(0.002)	(2.096)	(2.572)		
Truancy						
α_{21}	-0.008	-0.011^{*}	1.228	-0.529	195544	0.314
	(0.007)	(0.006)	(5.099)	(2.598)		
α_{22}	-0.002	-0.006^{*}	-1.670	0.751		
	(0.003)	(0.003)	(7.187)	(3.755)		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of the general deterrence effect (α_{21}) and the specific deterrence effect (α_{22}) in equation (2) for different categories of misbehaviors, estimated separately. Columns 2-5 show the results using different estimation methods. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include a control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. The sample for each category does not include schools with less than 5 students' first offense observations in the category or without second offense observations in the category in the academic year. The mean of the dependent variable is reported in the last column. Standard errors are reported in parentheses and clustered at the student level.

Table 9: General Deterrence Effects for the Second Offense by OSS experience

Offense Category	Sample with OSS Experience		Sample without OSS Experience	
	OLS & GFE	2SLS & GFE	OLS & GFE	2SLS & GFE
Violence	-0.054***	-0.117	-0.068***	-0.264
	(0.006)	(0.117)	(0.009)	(0.206)
	226697	226697	125132	125132
	0.260	0.260	0.310	0.310
Disrespect	-0.057***	-0.206	-0.121***	-0.348***
	(0.008)	(0.410)	(0.007)	(0.116)
	119209	119209	288115	288115
	0.404	0.404	0.425	0.425
Truancy	0.014	0.854	-0.022***	-1.231
	(0.013)	(0.883)	(0.008)	(0.909)
	33936	33936	161203	161203
	0.301	0.301	0.315	0.315

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{21} (general deterrence effects on students' second offense) in equation (2) by separately using the student-observations that were out-of-school suspended (OSS) for their first offense ($p_{1is\tau} = 1$) and the student-observations that were not out-of-school suspended for their first offense ($p_{1is\tau} = 0$). The third number in each cell is number of observations. The third number in each cell is the sample size. The fourth number in each cell is the mean of the dependent variable (the offense indicator). Standard errors are reported in parentheses and clustered at the student level.

Table 10: The Effects of Suspension Experience on Offending Students' Achievement Outcomes

Achievement Outcome	All Students	Only Offending Students						
		Ever OSS or Not in the Year					OSS/Not for 1st Offense	
	OLS Limited Controls	OLS Limited Controls	OLS Full Controls	OLS & GFE	2SLS	2SLS & GFE	OLS & GFE	2SLS & GFE
End of Grade Math (Grade 3-8)	-0.206***	-0.143***	-0.013***	-0.012***	0.068	0.051	-0.025***	0.125
	(0.003)	(0.003)	(0.002)	(0.002)	(0.187)	(0.178)	(0.003)	(0.293)
	3145909	548893	548893	548893	548893	548893	548893	548893
Dropout (Grade 9)	0.185***	0.154***	0.072***	0.050***	-0.077	-0.216	0.038***	-0.356
	(0.002)	(0.003)	(0.003)	(0.003)	(0.186)	(0.187)	(0.003)	(0.319)
	165046	114642	114642	114642	114642	114642	114642	114642
ACT Composite Score (Grade 9-10)	-1.112***	-0.999***	-0.211***	-0.178***	-2.741	-2.068	-0.120***	-5.956
	(0.026)	(0.029)	(0.024)	(0.024)	(1.824)	(1.807)	(0.029)	(5.979)
	115719	97419	97419	97419	97419	97419	97419	97419

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{31} in equation (3) – the effects of out-of-school suspension on offending students' achievement outcomes. The achievement outcomes include end-of-grade test math score (normalized), a dropout indicator (finally graduate or dropout from the high school) and ACT composite score (points). The estimation for dropout and ACT composite scores uses suspension experience of students in grade 9 and students in grade 9-10 respectively. While the second column shows the results using the "all students" sample, other columns show the results by limiting the sample to students with an offense record – students at risk of suspension. The key explanatory variable for Columns 2-6 is an indicator of whether the student was ever suspended in the academic year. For Columns 7-8, it is an indicator of whether the student was suspended for the first offense in the academic year. OLS (Limited Controls): a model with only school-level control variables and school fixed effects; OLS (Full Controls): a model with all control variables (student-level and school-level) and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include the control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. Standard errors are reported in parentheses and clustered at the student level. The third number in each cell is the sample size.

Table 11: The Effects of Suspension Experience on Achievement by Race

Achievement Outcome by Race	OLS	OLS & GFE	2SLS	2SLS & GFE	Sample Size
End of Grade Math Test Score					
White	-0.018***	-0.017***	-0.328	-0.242	212389
(Grade 3-8)	(0.003)	(0.003)	(0.702)	(0.526)	
Black	-0.008**	-0.008**	0.300	0.305	241130
(Grade 3-8)	(0.003)	(0.003)	(0.223)	(0.227)	
Dropout					
White	0.074***	0.050***	-0.289	-0.383	47751
(Grade 9)	(0.005)	(0.005)	(0.422)	(0.413)	
Black	0.066***	0.046***	0.113	-0.033	47750
(Grade 9)	(0.004)	(0.004)	(0.260)	(0.249)	
ACT Composite Score					
White	-0.205***	-0.178***	-3.345*	-2.886	43385
(Grade 9-10)	(0.042)	(0.042)	(1.813)	(1.839)	
Black	-0.180***	-0.149***	0.029	1.565	38894
(Grade 9-10)	(0.033)	(0.033)	(6.906)	(6.008)	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{31} in equation (3) for white and black students – the effects of out-of-school suspension experience on white or black offending students' achievement outcomes. The achievement outcomes include end-of-grade test math score (normalized), a dropout indicator (finally graduate or dropout from the high school) and an ACT composite score (points). The estimation for dropout and ACT composite scores uses suspension experience of students in grade 9 and students in grade 9-10 respectively. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include the control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. Standard errors are reported in parentheses and clustered at the student level.

Table 12: The Overall Effects of Discipline on Students' Achievement Outcomes

Achievement Outcome by Student Group	OLS	OLS & GFE	2SLS	2SLS & GFE	Sample Size
End of Grade Math Score (Grade 3-8)					
All Students	0.003	-0.003	-0.014	-0.029	2711053
	(0.003)	(0.003)	(0.054)	(0.050)	
Students	-0.001	-0.003	0.103*	0.074	1225598
with no offense record	(0.003)	(0.003)	(0.061)	(0.055)	
End of Grade Math Score (Grade 6-8)					
All Students	0.002	0.001	0.168	0.116	1433766
	(0.003)	(0.003)	(0.126)	(0.111)	
Students	-0.001	0.002	0.222**	0.202**	649423
with no offense record	(0.004)	(0.004)	(0.105)	(0.101)	
Dropout					
All Students	0.003	0.001	0.068	-0.041	400444
(Grade 9)	(0.004)	(0.003)	(0.087)	(0.084)	
Students	0.005	0.004	0.019	0.050	153684
with no offense record	(0.003)	(0.003)	(0.049)	(0.045)	
ACT Composite Score					
All Students	0.025	0.076*	-0.545	-0.138	444065
(Grade 9-10)	(0.046)	(0.046)	(0.988)	(1.046)	
Students	-0.025	-0.008	-1.320	-1.394	197375
with no offense record	(0.067)	(0.067)	(1.008)	(1.042)	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{41} in equation (4) – the effects of out-of-school suspension on "never offending" students or all students' achievement outcomes. The achievement outcomes include an end-of-grade test math score (normalized), a dropout indicator (finally graduate or dropout from the high school) and an ACT composite score (points). The estimation for dropout and ACT composite scores uses suspension experience of students in grade 9 and students in grade 9-10 respectively. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include the control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. Standard errors are reported in parentheses and clustered at the student level.

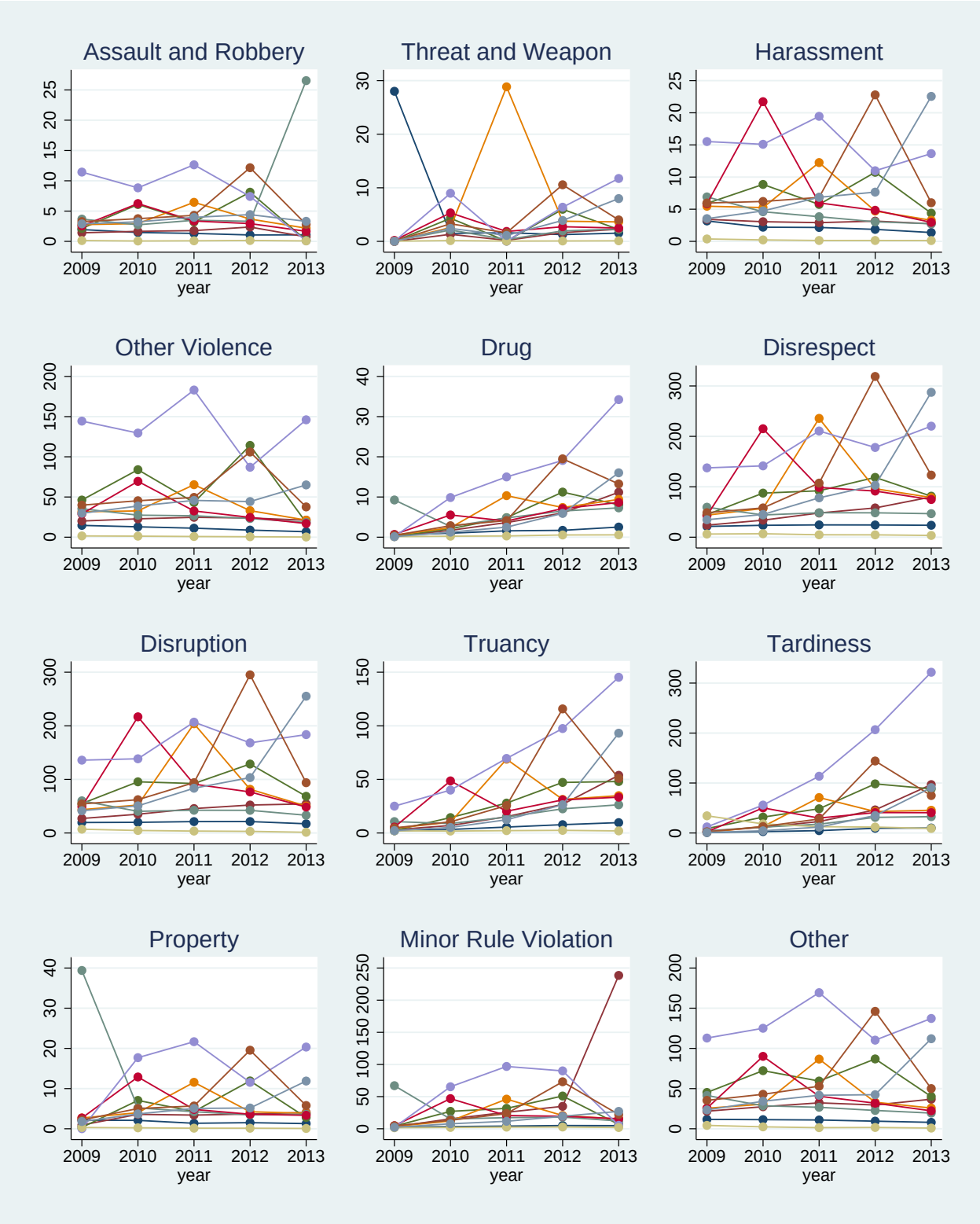
Table 13: The Overall Effects of Discipline on Achievement by Race

Achievement Outcome	White Students		Black Students	
	OLS & GFE	2SLS & GFE	OLS & GFE	2SLS & GFE
End of Grade Math Test Score				
(Grade 3-8)	0.000	0.084	0.004	-0.239
	(0.003)	(0.054)	(0.005)	(0.153)
	1420134	1420134	735245	735245
(Grade 6-8)	0.002	0.282***	0.004	2.058
	(0.004)	(0.099)	(0.007)	(1.891)
	765241	765241	392171	392171
Dropout				
(Grade 9)	-0.003	-0.017	0.007	0.107
	(0.004)	(0.060)	(0.007)	(0.344)
	217395	217395	114156	114156
ACT Composite Score				
(Grade 9-10)	0.147**	-1.059	-0.142*	1.343
	(0.063)	(0.799)	(0.074)	(2.254)
	256441	256441	115083	115083

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{41} for black and white students, estimated separately. The second and third columns show the results for white students; the fourth and fifth columns show the results for black students. The estimation for dropout and ACT composite scores uses the suspension experience of students in grade 9 and students in grade 9-10 respectively. OLS&GFE: a model with all control variables, school fixed effects and student unobserved heterogeneity; 2SLS&GFE: additionally instruments the DPI variable. The third number in each cell is sample size. Standard errors are reported in parentheses and clustered at the student level.

Figure 1: Estimated Student Group and Offense Numbers



Note: This figure shows the frequencies (in units of hundreds) of each category of offenses across years for each estimated group of students. The ten colors represent ten different estimated groups of students.

References

Aizer, Anna and Joseph J.Jr. Doyle (2015). "Juvenile incarceration, human capital and future crime: evidence from randomly-assigned judges." *The Quarterly Journal of Economics* 130 (2), pp. 759–803.

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

Becker, Gary (1968). "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2), pp. 169–217.

Bonhomme, Stéphane and Elena Manresa (2015). "Grouped patterns of heterogeneity in panel data." *Econometrica* 83 (3), pp. 1147–1184.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2016a). "A Distributional Framework for Matched Employer Employeee Data." *Working Paper*.

— (2016b). "Discretizing Unobserved Heterogeneity: Approximate Clustering Methods for Dimension Reduction." *Working Paper*.

Doyle, Joseph J.Jr. (2007). "Child protection and child outcomes: Measuring the effects of foster care." *The American Economic Review* 97 (5), pp. 1583–1610.

Dreikurs, Rudolf (1968). *Psychology in the classroom: a manual for teachers*. New York, Harper & Row.

Ehrlich, Isaac (1973). "Participation in illegitimate activities: A theoretical and empirical investigation." *The Journal of Political Economy* 81, pp. 521–565.

Forgy, Edward W (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications." *Biometrics* 21, pp. 768–769.

Gregory, Anne, Russell J Skiba, and Pedro A Noguera (2010). "The achievement gap and the discipline gap two sides of the same coin?" *Educational Researcher* 39 (1), pp. 59–68.

Heckman, James J and Edward Vytlacil (2005). "Structural equations, treatment effects, and econometric policy evaluation1." *Econometrica* 73 (3), pp. 669–738.

Imbens, Guido W and Joshua D Angrist (1994). "Identification and estimation of local average treatment effects." *Econometrica* 62 (2), pp. 467–475.

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

Kling, Jeffrey R (2006). "Incarceration length, employment, and earnings." *The American economic review* 96 (3), pp. 863–876.

Lee, Talisha, Dewey Cornell, Anne Gregory, and Xitao Fan (2011). "High suspension schools and dropout rates for black and white students." *Education and Treatment of Children* 34 (2), pp. 167–192.

Lin, Chang-Ching and Serena Ng (2012). "Estimation of panel data models with parameter heterogeneity when group membership is unknown." *Journal of Econometric Methods* 1 (1), pp. 42–55.

Losen, Daniel J and Russell J Skiba (2010). "Suspended education: Urban middle schools in crisis." *Southern Poverty Law Center*.

Morris, Edward W and Brea L Perry (2016). "The Punishment Gap: School Suspension and Racial Disparities in Achievement." *Social Problems* 63 (1), pp. 68–86.

Perry, Brea L and Edward W Morris (2014). "Suspending progress collateral consequences of exclusionary punishment in public schools." *American Sociological Review* 79 (6), pp. 1067–1087.

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

Skiba, RJ and MK Rauch (2015). "Reconsidering exclusionary discipline: The efficacy and equity of out-of-school suspension and expulsion." *Handbook of Classroom Management*, pp. 116–138.

Skiba, Russell J and Kimberly Knesting (2001). "Zero tolerance, zero evidence: An analysis of school disciplinary practice." *New Directions for Youth Development* 2001 (92), pp. 17–43.

Steinley, Douglas (2006). "K-means clustering: a half-century synthesis." *British Journal of Mathematical and Statistical Psychology* 59 (1), pp. 1–34.

Stock, James H and Motohiro Yogo (2005). "Testing for weak instruments in linear IV regression." *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80–108.

Wald, Johanna and Daniel J Losen (2003). "Defining and redirecting a school-to-prison pipeline." *New Directions for Youth Development* 2003 (99), pp. 9–15.

Wettach, Jane, Jenni Owen, and Claire Katie Hoffman (2015). "Instead of Suspension: Alternative Strategies for Effective School Discipline." *Duke Center for Child and Family Policy and Duke Law School*.

A

Offense Categories and Types

Violence

1.Assault, Robbery and sexual offense

001 = Assault resulting in a serious injury
002 = Assault involving the use of a weapon
104 = Physical attack with a firearm or explosive device
003 = Assault on school personnel not resulting in a serious injury
071 = Assault on non-student w/o weapon and not resulting in serious injury
044 = Assault on student
072 = Assault on student w/o weapon and not resulting in serious injury
090 = Violent assault not resulting in serious injury
045 = Assault - other
103 = Robbery with a firearm or explosive device
010 = Robbery with a dangerous weapon
093 = Robbery without a weapon
016 = Kidnapping
023 = Extortion
015 = Taking indecent liberties with a minor
012 = Rape
013 = Sexual offense
014 = Sexual assault not involving rape or sexual offense

2.Threat and possession of a weapon

043 = Bomb threat
105 = Threat of physical attack with a firearm
106 = Threat of physical attack with a weapon
107 = Threat of physical attack without a weapon
019 = Communicating threats (G.S. 14-277.1)
008 = Possession of a firearm or powerful explosive
009 = Possession of a weapon (excluding firearms and explosives)

3.Disorderly conduct and harassment

022 = Disorderly conduct (G.S. 14-288.4(a)(6))
038 = Harassment - sexual
101 = Harassment - Racial
102 = Harassment - Disability
109 = Harassment âĂŞ Sexual orientation
110 = Harassment âĂŞ Religious affiliation
025 = Harassment - verbal
080 = Discrimination

4.Other violent behavior

027 = Aggressive behavior
021 = Affray (G.S. 14-33)
024 = Fighting
026 = Hazing
052 = Bullying
094 = Cyber-bullying
079 = Gang activity

Drug

054 = Sale of controlled substance in violation of law - cocaine
055 = Sale of controlled substance in violation of law - marijuana
056 = Sale of controlled substance in violation of law - Ritalin
057 = Sale of controlled substance in violation of law - other
088 = Distribution of a prescription drug
049 = Use of controlled substances
048 = Use of alcoholic beverages
050 = Use of narcotics
070 = Use of tobacco

096 = Under the influence of controlled substances
095 = Under the influence of alcohol
005 = Possession of controlled substance in violation of law - cocaine
006 = Possession of controlled substance in violation of law - marijuana
007 = Possession of controlled substance in violation of law - Ritalin
017 = Possession of controlled substance in violation of law - other
020 = Alcohol Possession (G.S. 18B)
041 = Possession of tobacco
051 = Possession of chemical or drug paraphernalia
086 = Possession of student's own prescription drug
087 = Possession of another person's prescription drug

Disrespect

061 = UB: Disrespect of faculty/staff
032 = UB: Inappropriate language/disrespect
033 = UB: Insubordination

Property

053 = RO: Burning of a school building (G.S. 14-60)
036 = UB: Theft
039 = UB: Property damage
018 = UB: Unlawfully setting a fire (G.S. 14-277.1)

Truancy

075 = UB: Skipping school
030 = UB: Truancy
067 = UB: Leaving school without permission
074 = UB: Skipping class
066 = UB: Leaving class without permission

Other

1.Disruptive

042 = UB: Disruptive Behavior

2.Tardiness

064 = UB: Excessive tardiness
078 = UB: Late to class

3.Some other minor rule violations

028 = UB: Honor code violation
031 = UB: Dress code violation
060 = UB: Cell phone use
063 = UB: Excessive display of affection
068 = UB: Mutual sexual contact between two students
029 = UB: False fire alarm
035 = UB: Falsification of information
034 = UB: Gambling
059 = UB: Being in an unauthorized area
091 = UB: Misuse of school technology
040 = UB: Inappropriate items on school property
047 = UB: Use of counterfeit items
046 = UB: Possession of counterfeit items
065 = UB: No Immunization

4.Other

037 = UB: Bus misbehavior
077 = UB: Physical exam
114 = UB: Inappropriate Behavior
092 = UB: Repeat offender
058 = UB: Other School Defined Offense
069 = UB: Other

Table A1: General and Specific Deterrence Effects for the Second Offense

Offense Category (or Type)	OLS	OLS &GFE	2SLS	2SLS &GFE	Sample Size	Mean of Dependent Var.
Property						
α_{21}	-0.005	-0.003	-0.016	0.005	24420	0.13
	(0.009)	(0.009)	(0.858)	(0.560)		
α_{22}	0.016***	0.014**	-1.307	-0.922		
	(0.005)	(0.005)	(2.767)	(1.684)		
Drug						
α_{21}	-0.009	-0.009	0.224	0.111	36487	0.2
	(0.010)	(0.010)	(0.584)	(0.426)		
α_{22}	-0.011*	-0.009*	-0.060	-0.123		
	(0.005)	(0.005)	(0.431)	(0.351)		
Fighting						
α_{21}	-0.023***	-0.024***	-0.282	-0.175	133341	0.16
	(0.006)	(0.006)	(3.387)	(2.858)		
α_{22}	-0.013***	-0.016***	4.770	4.435		
	(0.004)	(0.003)	(8.614)	(7.192)		
Aggressive Behavior						
α_{21}	-0.017***	-0.021***	-1.121	-1.478	117886	0.22
	(0.006)	(0.006)	(2.822)	(3.390)		
α_{22}	0.005*	0.001	1.166	1.552		
	(0.003)	(0.003)	(3.094)	(3.715)		
Disrespect to Faculty						
α_{21}	-0.030***	-0.031***	1.043	0.650	122166	0.28
	(0.006)	(0.006)	(1.279)	(0.854)		
α_{22}	0.002	-0.004	-1.265	-0.797		
	(0.003)	(0.003)	(1.507)	(1.017)		
Insubordination						
α_{21}	-0.081***	-0.084***	0.380	0.474	224651	0.37
	(0.006)	(0.006)	(0.440)	(0.451)		
α_{22}	0.002	-0.006**	-0.619*	-0.708**		
	(0.003)	(0.003)	(0.330)	(0.331)		
Inappropriate Language						
α_{21}	-0.011*	-0.015***	0.204	0.071	166167	0.26
	(0.005)	(0.005)	(0.357)	(0.286)		
α_{22}	0.007**	0.001	-0.801	-0.437		
	(0.003)	(0.003)	(0.645)	(0.514)		
Skipping Class						
α_{21}	-0.030***	-0.023***	-0.218	-0.211	124294	0.27
	(0.007)	(0.007)	(0.219)	(0.211)		
α_{22}	0.004	-0.003	0.339	0.181		
	(0.004)	(0.004)	(0.443)	(0.430)		
Disruptive Behavior						
α_{21}	-0.122***	-0.128***	-0.851**	-0.697**	358033	0.40
	(0.005)	(0.005)	(0.340)	(0.335)		
α_{22}	-0.024***	-0.030***	0.655	0.618		
	(0.002)	(0.002)	(0.502)	(0.493)		
Excessive Tardiness						
α_{21}	-0.075***	-0.069***	-14.748	43.990	111400	0.45
	(0.011)	(0.011)	(35.668)	(442.025)		
α_{22}	-0.018**	-0.022***	18.460	-56.236		
	(0.008)	(0.007)	(44.752)	(564.759)		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of the "general deterrence effect" (α_{21}) and "the specific deterrence effect" (α_{22}) in equation (2) for different categories (or types) of misbehaviors, estimated separately. Columns 2-5 show the results using different estimation methods. OLS: a model with all control variables and school fixed effects; OLS&GFE: adds an additional control for student unobserved heterogeneity; 2SLS: instruments the DPI variable but does not include a control for student unobserved heterogeneity; 2SLS&GFE: instruments the DPI variable and controls for student unobserved heterogeneity. The sample for each category does not include schools with less than 5 students' first offense observations of the category or without second offense observations of the category in the academic year. The mean of the dependent variable is reported in the last column. Standard errors are reported in parentheses and clustered at the student level.

Table A2: General Deterrence Effects for the Second Offense by OSS Experience

Offense	Sample with OSS Experience		Sample without OSS Experience	
Category (or Type)	OLS & GFE	2SLS & GFE	OLS & GFE	2SLS & GFE
Property	0.012	0.257	-0.017	-0.056
	(0.013)	(0.353)	(0.013)	(0.234)
	12927	12927	11493	11493
Drug	0.006	0.110	-0.055***	-0.247
	(0.013)	(0.749)	(0.018)	(0.452)
	20907	20907	15580	15580
Fighting	-0.025***	1.754**	-0.056***	298.6
	(0.007)	(0.777)	(0.014)	(39616.6)
	114579	114579	18762	18762
Aggressive Behavior	0.003	-0.282	-0.044***	-0.108
	(0.009)	(0.213)	(0.009)	(0.323)
	51729	51729	66157	66157
Disrespect to Faculty	0.013	0.263	-0.062***	-0.079
	(0.011)	(0.227)	(0.008)	(0.137)
	41403	41403	80763	80763
Insubordination	-0.009	-0.113	-0.127***	-0.423
	(0.011)	(0.251)	(0.008)	(0.297)
	61334	61334	163317	163317
Inappropriate Language	0.021**	-0.024	-0.034***	-0.184*
	(0.009)	(0.195)	(0.007)	(0.107)
	59099	59099	107100	107100
Skipping Class	0.036**	-0.212	-0.024***	0.087
	(0.017)	(0.309)	(0.009)	(0.141)
	17558	17558	107881	107881
Disruptive Behavior	-0.060***	-0.556	-0.152***	-0.498***
	(0.008)	(0.473)	(0.007)	(0.177)
	83497	83497	278704	278704
Excessive Tardiness	-0.115***	2.775	-0.062***	0.659*
	(0.033)	(11.33)	(0.015)	(0.344)
	9028	9028	102372	102372

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table reports estimates of α_{21} (general deterrence effects on students' second offense) in equation (2) by separately using the student-observations that were out-of-school suspended for their first offense ($p_{1is\tau} = 1$) and the student-observations that were not out-of-school suspended for their first offense ($p_{1is\tau} = 0$). Standard errors are reported in parentheses and clustered at the student level. The third number in each cell is number of observations.