The Changing Impact of School Suspensions on Student Outcomes: Evidence from North Carolina Public Schools

Lewis Zhu

Professor Jason Baron, Faculty Advisor
Professor Duncan Thomas, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University.

Duke University
Durham, North Carolina
2025

Acknowledgements

I would like to express my sincere gratitude to Professor Duncan Thomas for his extraordinary mentorship, encouragement, and kindness over the past two and a half years. From research guidance to career advice and life conversations, his support has had a lasting impact on me. I feel incredibly fortunate to have had the opportunity to learn from him both in and outside of the classroom.

I am deeply grateful to Professor Jason Baron for his thoughtful advising and consistent support. Since first taking his class two semesters ago to the many conversations we've had since, he has shown me genuine kindness and a commitment to my development as a student and researcher. His investment in both me and this project has made a lasting impact on my academic growth.

I have also benefited greatly from conversations with members of the Frankenberg-Thomas Lab and the Duke Economic Analytics Laboratory, whose feedback and encouragement helped refine my ideas and analysis along the way. I am thankful to everyone who took the time to engage with my work and offer thoughtful insights.

This research was generously supported by funding from the Undergraduate Research Support Independent Study Grant and the Economics Department Research Support Grant. I am also grateful to the staff at the North Carolina Education Research Data Center for their assistance with data access and documentation.

Finally, I would like to thank my friends and family. Their encouragement and unwavering support have meant everything to me throughout this process.

Abstract

Motivated by a historic decline in standardized test scores among US students, this paper investigates whether exclusionary discipline—specifically out-of-school suspensions (OSS)—contributes to changes in academic performance. Drawing on administrative data from the North Carolina Education Research Data Center, which span a period marked by substantial discipline policy reform, I assess whether OSS rates are associated with school-level achievement and estimate the effect of OSS on individual student outcomes. I find that these relationships vary over time. As statewide suspension rates have declined in recent years, the negative association between school-level OSS rates and achievement has weakened, while the effect of OSS on individual student outcomes has grown more negative. One interpretation is that, as suspensions become less common, being suspended is a stronger negative signal for the child, possibly inducing stigma and differential treatment which worsen outcomes. Another possibility is that suspensions have become more targeted, such that those who are still suspended may have engaged in more serious misbehavior associated with worse outcomes. Meanwhile, school-level estimates may appear less negative as suspensions now target a smaller group of (on average) more disruptive students.

JEL classification: H75; I21; I24; I28

Keywords: Suspension; Discipline policy reform; Test scores; Student achievement

1 Introduction

Over the past decade, the decline in performance on nationally representative assessments among US school-age children has been alarming. On the National Assessment of Educational Progress (NAEP), for example, average math and reading scores among 4th and 8th graders have reached 20-to-30-year lows, erasing decades of educational progress in the span of a few years (U.S. Department of Education et al., 2024a, 2024b). While part of the decline is likely linked to the impact of COVID-19 on schools and students, the downturn in test scores predates the pandemic. Specifically, NAEP scores began falling in the early- to mid-2010s, coinciding with sharp declines in school suspension rates between 2011 and 2014. For 8th grade math in particular, the steepest drop in performance occurred the year following the most pronounced decline in suspensions. The goal of this paper is to investigate the links between test scores and changes in school suspensions.

This is important because associations between low test scores in elementary and secondary school and poor outcomes later in life are well-documented. Student achievement is a strong predictor of college attendance and completion, future income, and even involvement with the criminal justice system (Chetty et al., 2014; Cook & Kang, 2016; Doty et al., 2022; Lochner, 2020). It is, therefore, critical to identify modifiable policy-based pathways that likely contribute to declines in student academic performance.

While the urgency of understanding these declines is clear, so too is the complexity of the task. A large existing literature has explored a wide array of potential factors that contribute to academic achievement, ranging from shifts in family characteristics and school funding to explanations surrounding teacher quality and classroom sizes. Section 2 reviews the literature that is key for this research. Many of these factors, however, are difficult to change or costly to

address, if not both. One angle that remains sparsely explored is that of school discipline, particularly the use of punitive measures that remove disruptive students from the school (i.e. exclusionary discipline), such as out-of-school suspensions (OSSs). Given that disciplinary practices are an aspect of school climate that is potentially amenable to manipulation by policy, and that there have been important policy changes in recent years, this is an ideal time to investigate the links with student performance.

On the national level, overall rates of OSS increased from 4% in 1973 to 7% by the 2009-2010 school year, with the mid-1980s and early-1990s seeing a particularly marked increase (Leung-Gagné et al., 2022). This rapid rise in suspension rates coincides with the adoption of several zero-tolerance policies targeting drug use and gun possession, most notably the Guns-Free Schools Act of 1994 which mandated a minimum one-year expulsion of students who brought firearms to school (American Psychological Association Zero Tolerance Task Force, 2008; Cerrone, 1999). As a philosophy, zero-tolerance uses punitive measures as a deterrent for disruptive behaviors (Ewing, 2000). To this end, such policies call for enforcement of predetermined consequences, typically severe and exclusionary, for specific offenses without consideration of potentially mitigating contexts. Eventually, schools began applying zero-tolerance to a broader range of behaviors deemed disruptive.

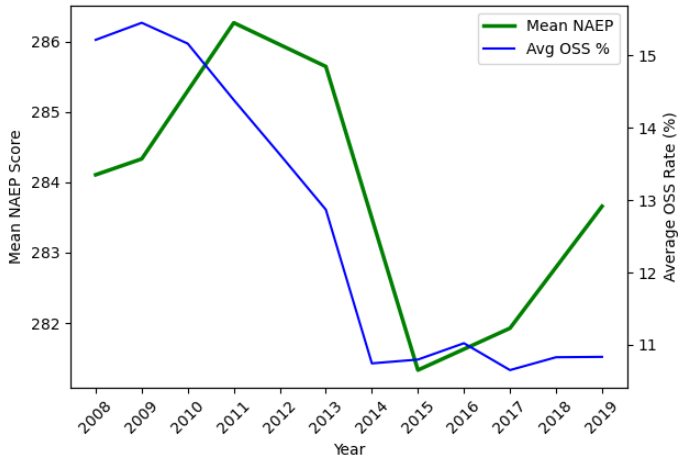
In the years following the implementation of these policies, a substantial body of research emerged suggesting adverse consequences. Numerous studies find a range of negative student academic and behavioral outcomes (Arcia, 2006; Noltemeyer et al., 2015), while others highlight the disproportionate impact of zero-tolerance practices on minority students, economically disadvantaged students, and students with disabilities, often exacerbating pre-existing educational disparities (Hoffman, 2014; Skiba & Peterson, 1999). By the early 2010s, such

evidence prompted policymakers and educators to reconsider the efficacy of zero-tolerance approaches, leading to the introduction of reforms aimed at limiting their use. In particular, policy efforts over the last decade have increasingly shifted toward promoting inclusionary disciplinary practices such as in-school suspensions (ISS) as an alternative to exclusionary measures such as OSS. However, the research underpinning these reforms is largely correlational and does not rule out the possibility that students who are suspended are also those who would have struggled academically or behaviorally regardless. Another possibility is that the observed relationships reflect reverse causality rather than the effects of suspension itself, raising the possibility that the resulting policy changes may be poorly targeted.

Between 2011 and 2014, over half of US states enacted legislation to curb the use of OSS and expulsions, yielding a nearly 20% decrease in OSS incidence nationally during that period (Council of State Governments Justice Center, 2017). At the federal level, the Obama administration advanced these efforts through the release of a comprehensive guidance package in 2014. The package provided states and districts with resources to reduce reliance on OSS and expulsions, promote safe and inclusive school environments, and ensure that disciplinary practices were compliant with federal civil rights laws, particularly addressing the disproportionate impact on students of color and students with disabilities (U.S. Department of Justice & U.S. Department of Education, 2014).

However, whether these reforms have yielded the intended benefits, particularly with respect to student achievement, remains an open question. As shown in Figure 1, the decline in suspension rates during the early 2010s coincided with a noticeable drop in Grade 8 NAEP math scores. While these parallel trends may appear suggestive, the two series diverge in later years,

Figure 1. Mean NAEP Grade 8 Math Score and OSS Rate by Grade (2008–2019)



Notes. Plot of OSS rate (calculated here as the mean of a student-level OSS receipt indicator variable across all students) within each school year on the right vertical axis. Note that NCERDC data have poor coverage of students pre-grade 3, so this OSS rate is the mean for grade 3–12 students. Average NAEP scores for the grade 8 math assessment are plotted on the left vertical axis.

and one cannot draw causal conclusions from these patterns alone. Nonetheless, the alignment in trends during the period of most pronounced disciplinary change offers motivation for a deeper examination. Understanding whether and how reductions in school suspension rates influence academic outcomes at the school level is therefore crucial for evaluating the broader impacts of these disciplinary reforms.

A growing body of research has sought to quantify the academic effects of exclusionary discipline. Yet, to my knowledge, relatively few papers are able to convincingly address the inherent endogeneity of suspension decisions. As discussed in the literature review in Section 2, a clear consensus on the magnitude or even direction of the effect of OSS on academic performance does not exist.

In this study, I draw on rich longitudinal education data maintained by the North Carolina Education Research Data Center (NCERDC). This restricted-use administrative dataset includes detailed student- and school-level data over multiple school years, allowing me to construct a comprehensive panel of students and schools. Importantly, the NCERDC data feature extremely high tracking rates of students across years—students can be followed throughout their

educational careers unless they drop out of the public school system or transfer to a private institution. Moreover, these data include information on the schools attended by each student, allowing me to link the student- and school-level data when necessary.

These detailed longitudinal data allow me to overcome traditional challenges in estimating the effects of OSS by employing a fixed effects (FE) strategy, exploiting within-unit changes over time to account for unobserved, time-invariant characteristics. In this paper, I implement this strategy across two complementary models. The first uses school-level data to examine the relationship between a school's OSS rate and its average academic performance within a given grade and year, relative to the statewide average. This model includes school FEs, which control for stable characteristics such as baseline levels of disorder, leadership style, or community socioeconomic context. The second model uses student-level data to estimate the effect of exclusionary discipline on individual academic outcomes, incorporating student FEs to account for time-invariant factors including innate ability, home environment, or persistent behavioral patterns. By including unit-level FEs, my approach may mitigate concerns that unobservable factors bias the estimated relationship between OSS and academic performance, allowing me to credibly argue that the variation in OSS rates/incidence is unrelated to remaining unit-level heterogeneity.

Ex ante, there are plausible mechanisms by which exclusionary discipline practices could lead to either positive or negative impacts on student achievement, at both the school and student level. At the school level, higher OSS rates could reduce overall academic performance if they reflect a punitive environment that disrupts learning or if they result in frequent instructional interruptions across classrooms. Conversely, they could improve academic outcomes if suspensions remove persistently disruptive students, thereby improving peer learning conditions

or reinforcing behavioral norms. At the student level, the direct effect of receiving a suspension may be negative, due to lost instructional time, stigmatization, or disengagement from school. However, it is also possible that suspension could generate an "effort effect" by deincentivizing future misbehavior and motivating improved academic focus for that student. Thus, the relationship between OSS and academic achievement is theoretically ambiguous at both the school and student level, underscoring the importance of rigorous empirical analysis tailored to each context.

In this paper, I find that higher school-level OSS rates are significantly associated with lower average test scores, even after accounting for school characteristics and time trends. However, this relationship weakens over time, particularly for math. At the individual level, receiving an OSS is associated with a modest but significant decline in test performance, even after controlling for student, school, and year FEs, and these effects become more negative in later years. This divergence suggests that the academic consequences of OSS are context-dependent and may have grown more severe as overall suspension rates have declined.

This paper contributes to two distinct but related literatures. First, it adds to the extensive body of work investigating potential explanations for national declines in test scores, such as those documented in the NAEP. A primary contribution of this paper is its focus on exclusionary discipline, a tractable and policy-relevant dimension of school climate. To this end, I examine the aggregate relationship between a school's OSS rate and its average academic performance within a given grade and year, relative to the corresponding state-wide average. While the analysis is descriptive, it helps to assess whether changes in exclusionary discipline are associated with recent test score declines, and whether this association varies with time.

The second contribution of this paper is to the literature examining the effect of exclusionary discipline on individual academic outcomes. While the school-level analysis offers a useful descriptive lens on broader trends, it necessarily masks heterogeneity in how students within those schools are affected. The second half of my analysis shifts focus to the student level, where I assess whether individual exposure to exclusionary discipline is associated with differences in test scores. Causal identification in this setting is challenging. I present arguments regarding the plausibility of a causal interpretation.

A key feature of this analysis is its use of more recent data from a period in which OSS rates have declined substantially. Whereas most existing studies use data from the early-2010s or earlier, when zero-tolerance policies were still widespread, my analysis takes advantage of the variation in suspension rates throughout the 2010s to ask whether the academic impact of OSS has changed as these policies have become less prevalent. In doing so, I introduce a new way of thinking about these effects, drawing inspiration from the marginal treatment effect (MTE) framework: as schools become more or less likely to suspend students, does the academic impact of receiving an OSS change as well? I provide evidence that it does, and in doing so, contribute to a more nuanced understanding of the relationship between discipline and learning. This analysis builds on a newer era of disciplinary reform and complements earlier causal studies, which primarily examined settings dominated by zero-tolerance policies.

The remainder of the paper will proceed as follows. In Section 2, I provide an overview of the existing literatures to which this paper contributes most meaningfully. In Section 3, I give an overview of how and why students are suspended in North Carolina as well as which specific kinds of offenses appear in my analysis. Additionally, I discuss the test scores that I use to measure academic achievement. Section 4 describes the data sources that are used in my analysis

and how my analysis sample is constructed. It also presents descriptive statistics that provide context for the results to come. In Section 5, I detail the empirical strategy with which I will estimate the effects of OSS. Section 6 presents the results of my analysis. Section 7 concludes.

2 Literature Review

While my study is motivated by the recent decline in test scores, much of the existing literature explaining trends in student performance focuses on earlier periods of improvement, particularly the substantial gains in US test scores prior to the 2010s. In the literature that focuses on aggregate trends in student performance, a prominent line of inquiry examines whether changes in funding for education can account for the observed patterns in achievement. Perhaps counterintuitively, Hanushek et al. (1996) find that the significant increases in educational funding between the late 1960s and early 1990s were ineffective in improving educational outcomes. In particular, both SAT and NAEP scores showed no consistent improvement during this period. More recently, Hanushek and Woessmann (2017) examine evidence on the causal relationship between school expenditures and student outcomes, concluding that effects are generally small or nonexistent. Moreover, Grissmer et al. (2000) find that the gains in NAEP scores observed during the 1990s are difficult to attribute to changes in school resources.

These conclusions are not without controversy. Indeed, a substantial portion of the literature contends that spending increases can improve student outcomes. Lafortune et al. (2018) examine relative NAEP performance from 1990 to 2011 in low-income school districts through an event study design and conclude that funding reforms caused gradual but significant increases in student achievement. Jackson et al. (2014) similarly exploit plausibly exogenous shifts in school spending induced by the timing of reforms and find causal evidence that sustained

increases in per-pupil spending are associated with improved long-term outcomes, including academic performance. Extending this line of inquiry, Jackson et al. (2021) examine whether the widespread declines in academic performance following the Great Recession can be attributed to reductions in school spending. This focus closely aligns with the trends I investigate, as my study is similarly motivated by the stagnation and reversal of test score gains in the years after 2010. They use a shift-share instrument based on states' pre-recession reliance on state funding to isolate exogenous variation in post-recession spending reductions. Their results suggest that students in states more vulnerable to state-level fiscal shocks experienced declines in test scores and college enrollment, with disproportionately negative effects on low-income and minority students. Moreover, Baron (2022) leverages the quasi-random results of close elections in expenditure decisions in a dynamic regression discontinuity strategy to show that test score improvements are linked to increases in operational spending, but not to investment in capital. While school finance reforms are clearly of policy relevance, the lack of consensus—both regarding whether resource-based interventions are effective and which types of spending matter most—poses a challenge for policymakers seeking to design targeted and cost-effective interventions.

The literature on teacher quality and teacher value-added has produced widely cited results that suggest important levers through which education policy might influence student achievement. For example, Hanushek and Rivkin (2006) provide a review of studies that draw from this literature up to 1994 and find that commonly measured teacher characteristics, such as education level, salary, test scores, and experience, show mostly weak or inconsistent correlations with student outcomes. However, more recent work has found some meaningful effects. Studies by Chetty et al. (2014), Havard et al. (2018), and others demonstrate strong

positive impacts of teacher value-added on a range of student outcomes, including academic performance.

A closely related literature examines class size, offering another school-based factor that may influence test score trends. Project STAR, a large-scale randomized experiment conducted in Tennessee, assigned students in kindergarten through third grade to either small or large classes and found strong evidence that smaller class sizes were associated with lasting improvements in student performance (Finn & Achilles, 1990; Mosteller, 1995; Schanzenbach, 2006). Similar findings emerged from Wisconsin's SAGE initiative, which followed a comparable design (Ehrenberg et al., 2001; Molnar et al., 1999). However, the interpretation of these have been debated. Hanushek et al. (1996) argue that efforts to reduce class size through increased school expenditures have generally failed to produce systematic improvements in student outcomes. Additionally, Hanushek (1999) raises methodological concerns about the STAR study, including concerns about successful random assignment and the non-random selection of participating schools, which may call into question the validity of the results.

There also exists a broad literature investigating the link between non-school factors and academic achievement. Grissmer et al. (1994) find evidence between the 1970s and 1990s that changes in parental education, family size, family income, and maternal age at birth are associated with gains in academic achievement. A meta-analysis by Castro et al. (2015) highlights particularly strong associations between student performance and family involvement in students' activities. Early intervention studies, such as those on Head Start and preschool programs, suggest long-term gains under certain conditions (e.g., Ludwig & Miller, 2007), though findings are mixed (Currie & Thomas, 1993; Garces et al., 2002; Gibbs et al., 2011; Lee et al., 1990; Ludwig & Miller, 2007; Zhai et al., 2012). A parallel literature in psychology links

test scores to traits like self-regulation and motivation (Duckworth et al., 2019), though such factors are difficult to influence at scale through policy.

Taken together, many of the factors discussed above are either costly to implement, difficult to influence through policy, or produce mixed findings that limit their practical relevance. Moreover, much of the existing literature—even recent work—relies on older datasets that may not reflect current policy environments or student populations. Only in recent years has data become available at the scale and granularity needed to examine newer potential contributors to test score trends. My study adds to this literature by focusing on a factor—disciplinary policy—that is both directly reformable and relatively low-cost, while also leveraging more recent data from a period of significant policy change.

Studies that estimate the causal effect of OSS on student educational outcomes form a small but important literature that is highly relevant to the question this paper investigates. This work uses primarily student-level data to estimate the impact of receiving an OSS on academic performance and related educational outcomes, with designs that aim to yield causal conclusions.

Two of these studies examine the same policy setting. Steinberg and Lacoe (2018) and Lacoe and Steinberg (2019) both study a disciplinary reform implemented by the School District of Philadelphia in the 2012–13 school year, which prohibited the use of OSS as punishment for two nonviolent infractions: failure to follow classroom rules or causing disruptions, and the use of profane or obscene language or gestures. Lacoe and Steinberg (2019) employ student-level panel data to control for time-invariant student FEs. This is complemented by an instrumental variable strategy which leverages variation in OSS receipt probability induced by the reform. Specifically, they instrument for OSS receipt using an indicator for whether a student committed one of the targeted infractions in the post-policy period. While the instrument satisfies the

relevance condition, the exclusion restriction is more difficult to justify. Exclusion requires that the policy affects student outcomes only through its effect on OSS receipt. This rules out, for instance, the possibility that the reform changed teacher behavior, school climate, or administrative responses in ways that also influence academic outcomes, independent of whether a student was suspended. Given the multifaceted nature of school reforms, this is not a plausible assumption. The study also has missing data concerns: in the post-policy period, infractions are only observed if they result in an OSS, requiring strong assumptions about the distribution of unobserved offenses. The authors conclude that OSS reduces achievement in both math and English, with small negative spillover effects on peer performance. While the methodological framework is appealing, the strength of the conclusions is weakened by these data and identification concerns.

Steinberg and Lacoe (2018) builds on this analysis by implementing a difference-in-differences strategy in the same setting. Their design compares students who received an OSS in the pre-reform period for infractions later targeted by the policy to students who did not, before and after the reform. Because the policy was district-wide and implemented at the same time for all schools, their approach does not rely on variation in exposure across treated and untreated schools. This nonstandard implementation complicates the interpretation of the difference-in-differences estimates, as there is no traditional untreated comparison group. The authors attempt to address this by conducting several robustness checks, including tests for parallel pre-trends in achievement across student groups. They find no significant overall effects of the reform on either treated students or their peers, though they do report negative peer effects in schools that failed to fully implement the policy.

Moreover, Hwang and Domina (2021) study grade 7–11 students in a California school district between the 2009-10 and 2011-12 school years. They attempt to identify the causal effect of suspension on peers by leveraging student and classroom FEs, which they claim plausibly sweep out all unobserved heterogeneity related to their outcome by controlling for fixed characteristics of students and classrooms. However, the authors devote limited attention to potential violations of their identifying assumption. For example, if the composition of classrooms changes over time in unobserved ways, or if teachers respond to peer disruptions through changes in instruction or grading that also affect student achievement, the estimates may be biased. Despite these concerns, they report that student achievement increases when classmates are suspended, particularly when the suspensions are for disruptive behavior. This suggests that there may be positive academic spillovers from the removal of disruptive peers.

Anderson et al. (2017) examine the impact of the number of OSS days received by a student in the prior school year on their current academic outcomes using individual-level data from Arkansas between 2008 and 2013. This period includes substantial policy shifts, which make the paper's null to slightly positive findings on test scores particularly noteworthy. In contrast to prior research conducted in contexts with stricter zero-tolerance environments, these results raise the possibility that the academic consequences of OSS may depend on the broader disciplinary climate. However, the study reports only an average effect across all years, leaving open the question of whether the relationship between OSS and achievement varies over time. The identification strategy faces challenges similar to those in Hwang and Domina (2021), particularly regarding time-varying unobserved confounders. While the findings are suggestive, the strength of the causal interpretation remains limited.

Yaluma et al. (2022) provide a complementary perspective through a school-level analysis of data from Ohio in the 2011-12, 2013-14, and 2015-16 school years. Their identification strategy relies on school and year FEs. The results indicate that OSS is generally associated with lower academic proficiency rates at the school level, though for certain demographic groups, receiving multiple suspensions within a single year appears to have a positive effect. The authors refrain from making strong causal claims, acknowledging that their FE strategy does not fully address potential endogeneity. Notably, this study is, to my knowledge, the first to allow the effects of OSS to vary by year. However, this is not a central focus of the analysis, and the authors find little evidence of temporal heterogeneity. The study's limited time frame—just three school years—also constrains its ability to detect meaningful changes over time, particularly as it omits earlier periods when disciplinary practices were undergoing rapid transformation.

This literature offers valuable insights but remains limited in scope, and the results are not uniformly convincing. The findings on whether OSS harms the suspended student and the effects on peers are mixed, with some evidence suggesting potential academic benefits. In addition, most of the existing literature focuses on relatively narrow time windows, primarily during periods when zero-tolerance policies either prevalent or still in the process of being repealed. My paper complements this research by using more recent data, covering a period of substantial shifts in disciplinary policy and practice. This expanded timeframe allows me to explore how the academic consequences of OSS may have evolved as schools have moved away from more punitive disciplinary models.

3 Setting

In North Carolina, schools exercise considerable discretion when making suspension decisions, as outlined by state policies (Sorensen et al., 2022). For less serious incidents, administrators decide whether to suspend a student, the type of suspension (ISS or OSS), and the duration (up to 10 days). Principals may recommend to the superintendent that a student be removed from school for 11 to 365 days for more severe offenses.

Disciplinary incidents are first referred to a school administrator by a teacher or another school actor before any consequence is determined. Once a consequence is assigned, the incident and outcome may be recorded through a statewide administrative data reporting system. Offenses that lead to the student receiving ISS, OSS, expulsion, assignment to an alternative learning program or school, corporal punishment, or that fall under one of the twelve serious "reportable offenses" must be reported (North Carolina Department of Public Instruction, 2021). Any reported offense appears in my data. In particular, since OSS incidents are mandated to be reported regardless of offense type, my analysis of OSS effects is not subject to concerns about selective reporting introducing bias or endogeneity related to which kinds of OSS cases are captured in the data.

While the motivation for this analysis is rooted in concerning national trends in academic achievement as measured by the NAEP, I do not use these test scores in my empirical analysis. NAEP data, though useful for identifying broad correlations, are reported at a high level of aggregation and cannot be used to construct a panel suitable for limiting the impact of unobserved heterogeneity in my empirical models. The identification strategy I employ, described in more detail in Section 5, hinges on the availability of a consistent panel structure over time. Hence, to measure student achievement, I must use an assessment that is required for

all students to take. For this reason, I opt for End-of-Grade (EOG) test scores in math and reading which are administered to and mandatory for all grade 3–8 students[1] in North Carolina (North Carolina Department of Public Instruction, 2025).

The EOG tests were implemented in response to the No Child Left Behind (NCLB) Act of 2001, which aimed to increase school accountability for student achievement and reduce disparities in educational outcomes (Dee & Jacob, 2011). NCLB required that all public schools receiving federal funding administer annual standardized tests to students in grade 3–8. This requirement was later retained under the Every Student Succeeds Act (ESSA) in 2015 (U.S. Department of Education, 2025), ensuring the continued use of annual assessments.

4 Data Sources and Sample Construction

4.A Description of Administrative Data Sources

The data that I use for my analysis are housed at the North Carolina Education Research Data Center (NCERDC). They maintain a comprehensive collection of administrative datasets covering all public, charter, and magnet schools in North Carolina. The data are organized by school year, with a separate file for each year. Crucially, NCERDC employs encrypted student identifiers that allow researchers to link individual students across tests, grades, and schools over time. Similarly, standardized school and local education agency (LEA) codes enable the merging of student datasets with detailed school-level files. This longitudinal structure permits the construction of consistent panels at both the student and school levels, which is essential for the FE models used in the analysis. The ability to track students and schools over multiple years,

[1] With the exception of students with disabilities who participate in statewide testing by taking an alternative assessment. These students make up roughly 4% of my data.

combined with the near-universal coverage of public schools, makes the data particularly well-suited for studying within-unit changes over time. Note that reporting on private schools is extremely limited, and so our analysis will focus on schools that receive some form of public funding.

Despite the use of encrypted student identifiers, the NCERDC data contain detailed information that could potentially be identifying. Due to the sensitive nature of the data, public access is not permitted. Researchers seeking to use the data must submit a formal project proposal to the NCERDC, specifying the objectives of their research, detailing the specific datasets requested, and explaining why access to identified data is necessary to accomplish their research goals. Approval from the NCERDC is required before access is granted, ensuring that all data usage adheres to strict confidentiality and data protection standards.

Before proceeding, it is important to clarify the terminology used throughout this paper regarding the unit of observation. Specifically, unless stated otherwise, references to "student-level" data refer to the student-by-school-year level, while "school-level" data typically refer to the school-by-grade-by-school-year level. The reason for the school-level data to also include the grade dimension is to capture potential grade-level heterogeneity in the relationship between OSS rates and academic achievement. Although the raw datasets are organized separately by year at the individual student and school levels, I link them across years to construct consistent panels for analysis.

The student-level analysis draws primarily from two NCERDC datasets, the first of which is a yearly master build file containing one record per student per academic year. Each record includes detailed information on standardized test performance, specifying the test date, score, and type. For students in grades 3 through 8 who are not eligible for alternative

assessments, this corresponds to the EOG exams in math and reading. The dataset also records exemption status and provides an explanation in the rare instance that a test score is missing. In addition to test data, the master build file includes a range of student-level characteristics, such as the number of days absent, as well as demographic information including race, sex, and economically disadvantaged status. A small number of students appear multiple times per year in this dataset. These are typically the case if students transferred from one North Carolina public school to another during a school year. In these cases, I keep the one observation from the student for which we have an EOG score.

The second NCERDC dataset used in this analysis is the student suspension file, which provides detailed information on disciplinary actions, specifically OSS. Following a 2001 mandate by the North Carolina legislature, the state began systematically collecting and reporting annual data on student disciplinary consequences. This dataset contains a record for each instance in which a legally reportable offense occurs, including every OSS, referral to an alternative school or program, or expulsion administered during the school year. However, OSS reporting was inconsistent and incomplete between 2002 and 2007, making those years unsuitable for analysis. Beginning in 2008, OSS reporting improved substantially in both quality and coverage. Accordingly, my analysis focuses on the period from 2008 to 2019. This period is well-suited for the study, as it captures the tail end of the widespread adoption of zero-tolerance policies and extends through several years following the Obama administration's 2014 federal guidance aimed at reducing the use of OSS. OSS rates declined sharply after 2019 due to the COVID-19 pandemic, as the shift to remote learning limited the feasibility of enforcing suspensions. While suspension rates have rebounded in recent years, it remains difficult to disentangle pandemic-related disruptions from broader trends. Thus, the 2008–2019 window

provides the ideal timeframe for examining the relationship between OSS rates and academic outcomes.

The school-level data used in this analysis come primarily from the NCERDC's public school universe dataset, a North Carolina-specific subset of the National Center for Education Statistics (NCES) Public School Universe files. This dataset includes one record per public school per school year, providing information such as the school's address, locale code, total number of teachers, number of economically disadvantaged students, and student enrollment by grade, disaggregated by race and gender. Additional school-level variables used in the analysis—most notably the year and grade-specific OSS rates—are constructed directly from the student-level data described earlier by aggregating to the school-by-grade-by-year level.

4.B Sample Construction

To construct the key suspension measure used in my analysis, I define a student-level OSS indicator variable that equals one if a student received at least one OSS during a given school year, and zero otherwise. This variable serves as the primary exposure metric in the student-level analysis. For the school-level analysis, I calculate the OSS rate as the fraction of students in a given school, grade, and year who received at least one OSS. This aggregation yields a consistent and interpretable measure of suspension practices across schools and over time.

I apply a number of sample restrictions to the data that I analyze. First, for the student-level analysis, I restrict the sample to students who have at least one test score observed both before and after 2014. This ensures that included students contribute to the analysis of potential changes associated with the Obama administration's 2014 federal guidance on school discipline.

This restriction does not apply to the school-level analysis, which aggregates all available students in a given school, grade, and year regardless of their individual observation windows. Second, the school-level analysis focuses on middle school grades (6 through 8). Although my data include grades 3 through 8, I exclude grades 3–5 because suspension rates among elementary school students are generally low, and suspending very young children may have qualitatively different implications that are less aligned with the core questions of this study. For the student-level analysis, I retain all observations for students in grades 3 through 8. This allows me to maximize the number of repeated observations per student. Since the identification strategy relies on student FEs, a sufficient number of observations per student is crucial to obtain reliable within-student estimates. Finally, I exclude students who participate in statewide testing through alternative assessments, such as NCEXTEND1 or NCEXTEND2, which are typically administered to certain students with disabilities. These assessments are scored on a different scale from the standard EOG tests, making comparisons difficult. Approximately 4% of students fall into this category and are removed from the analysis to maintain consistency in the measurement of academic performance. Ultimately, my school- and student-level analyses are performed across 24,949 and 3,058,431 observations, respectively.

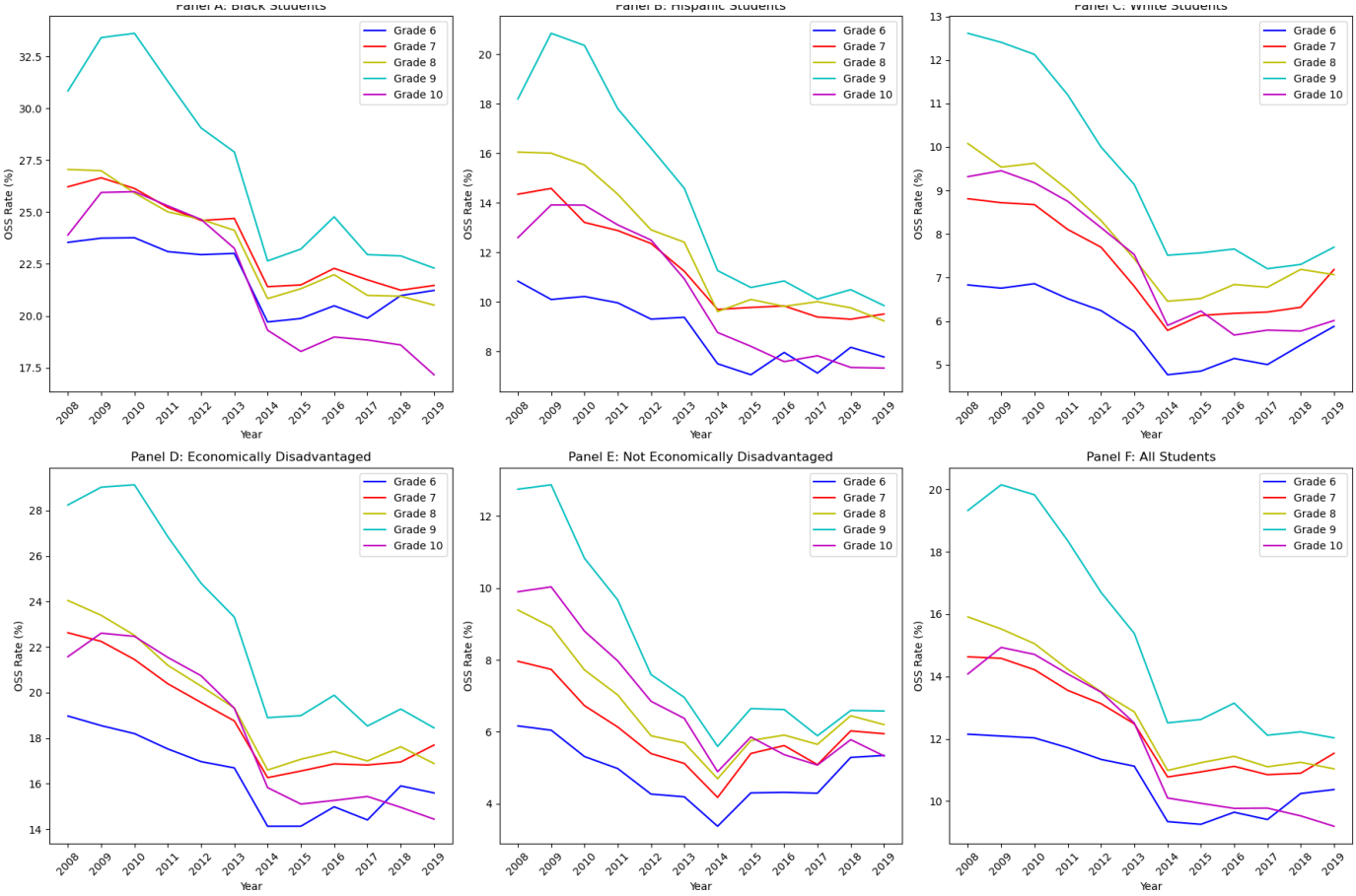
To ensure comparability of test scores over time, I standardize students' EOG scores within grade and school year. The raw EOG scale scores are not directly comparable across years, as the score ranges have been frequently adjusted. For example, reading scores ranged from 303 to 384 between 2008 and 2013, but shifted to a range of 406 to 488 from 2013 to 2019. Math scores experienced even more frequent changes in scaling. In addition, the score cutoffs for different achievement levels (e.g. proficiency) are determined by a committee of educators, and it is unclear whether these benchmarks are set before or after observing the distribution of scores.

Consequently, relying on rates of achievement levels, such as proficiency rates, as a measure of academic performance could lead to misleading inferences, since fluctuations in these rates may reflect shifting standards rather than actual changes in student performance. Standardizing the EOG scores eliminates these concerns by placing scores on a consistent scale within each grade and year. While this approach does not allow for interpretation of changes in statewide mean achievement over time, it enables meaningful comparisons of relative performance across schools and facilitates analysis of the relationship between OSS rates and academic outcomes relative to average state-wide performance within the same year. For our school-level analysis, we simply take the average EOG scores (now standardized) for each school, school year, and grade. For the rest of the paper, "EOG score" will refer to standardized (not scale) scores.

4.C Descriptive Statistics

To motivate our analysis, I begin by presenting some broad trends discussed in Section 1. Figure 1 (in Section 1) plots the average OSS rate over time alongside the national trend in Grade 8 NAEP math scores for the same period. Notably, the two series appear to share common trends and follow each other closely, with shifts in OSS rates corresponding to similar movements in NAEP performance. Additionally, the figure reflects the impact of policies implemented in the early 2010s aimed at curbing OSS, evidenced by the sharp decline in suspension rates. Interestingly, there is a noticeable flattening of OSS rates around 2014, despite that being the year in which the Obama administration issued its federal guidance discouraging exclusionary discipline practices. While these patterns are suggestive, they underscore the need for a more rigorous empirical investigation to better understand the relationship between OSS rates and student achievement.

Figure 2. OSS Rate by Grade by Demographic Group (2008–2019)



Notes. Panel A, B, and C plot OSS rate by grade (for grades 6–10) for Black, Hispanic, and White students, respectively. Panel D and E plot OSS rate by grade for economically disadvantaged students and non-disadvantaged students, respectively. Panel F shows overall OSS rate over time by grade. Note that the vertical axes of these six panels are different. This choice was made to illustrate the similarity of trends across groups. See Appendix 1 for the same plots on a shared axis, which better demonstrates the differences in OSS rate levels across demographic groups. The year on the horizontal axes denotes the year of the spring semester for each school year (e.g. 2013 denotes the 2012-13 school year).

Figure 2 illustrates how OSS rates have evolved over time across grades 6 through 10 across key demographic subgroups. The trends reveal similar temporal patterns across all groups, with a general decline in OSS rates beginning in the early 2010s and a flattening out by 2014. However, consistent with prior research, there are clear level differences between subgroups: Black and economically disadvantaged students are suspended at higher rates compared to their peers. These differences reflect well-documented disparities in disciplinary practices. Notably,

while the overall rates decline, the policy shifts during this period do not appear to significantly narrow the gaps between demographic groups.

Unlike the OSS data, presenting descriptive trends for EOG scores is less informative. Plotting raw scale scores over time is not meaningful, as the underlying scale ranges and proficiency cutoffs vary across years and are possibly determined subjectively by a committee of educators. This makes it unclear whether observed trends reflect actual changes in student achievement or shifting standards. Furthermore, because I standardize EOG scores within grade and school year, any time-series plot of the standardized scores at the state level would show no variation in means by construction. However, by *z*-scoring the EOG scores, I can meaningfully analyze variation in average test performance at the school level.

5 Empirical Strategy

An initial approach to analyzing the relationship between OSS rates and academic performance might involve regressing a school's average EOG scores. Even if a variety of covariates to account for observable school characteristics are included, it is unlikely to yield unbiased estimates of the causal effect of OSS rates on test scores. Schools with differing suspension rates may systematically differ along unobservable dimensions—such as leadership style, disciplinary philosophy, or broader school climate and culture—that also influence student achievement. Failing to account for these unobserved factors would confound the estimated relationship, making it difficult to disentangle the true impact of OSS rates from the influence of these underlying differences.

To address concerns about omitted variable bias, I leverage the rich longitudinal structure of the NCERDC data and implement a FEs strategy. I employ school FEs to control for time-

invariant unobservable characteristics at the school level, allowing me to capture baseline differences across schools. Additionally, I include year FEs to account for statewide shocks, policy changes, or other common factors affecting all schools in a given year. This approach allows me to focus on within-school variation over time, effectively comparing each school to itself as its OSS rate changes, while controlling for year-specific trends.

Let $OSSRate_{sgt}$ be the OSS rate within school s, grade g, and time period t. We consider specifications with both $t \in \{2008, \ldots, 2019\}$, and $t \in \{\text{Pre-2014}, \text{Post-2014}\}$. I estimate the following regression specification:

$$y_{sgt} = \beta_t OSSRate_{sgt} + \lambda_s + \delta_t + \gamma W_{sgt} + u_{sgt}, \quad (1)$$

where y_{sgt} is the average EOG scores for the math or reading assessments, λ_s and δ_t are school and year FEs, respectively, W_{sgt} is a vector of controls, and u_{sgt} is all remaining unobserved determinants of y_{sgt}. Controls that I encode in W_{sgt} include racial compositions of schools, proportion of students who are economically disadvantaged, and school year and grade indicators. Standard errors are clustered at the school level to account for the mechanical correlation in outcomes due to the same school appearing multiple times in the dataset.

For each of my outcomes, the coefficient of interest is β_t. By allowing β_t to vary with time in my specification, it enables me to explore heterogeneity in estimated effects of OSS rate on test scores by year (and by pre- versus post-2014, which is the year of the Obama administration guidance to curtail the use of OSS), which specifically facilitates an exploration of how the relationship evolves over time. This is a central component of my contribution to the literature, as it provides a framework for examining whether changes in OSS rates over the study period were accompanied by shifts in their impact on academic achievement. Section 6 also includes estimates of the average β_t over all t which I denote as β. This allows me to document

the importance of school and year FE and additional time-varying controls. I also present heterogeneity by school types, although I do not do this in my primary specification.

Estimates of β_t cannot be interpreted as causal since the necessary condition $E\left[OSSRate_{sgt} u_{sgt} \mid \lambda_s, \delta_t \right] = 0$ likely fails. In other words, even though we condition on time-invariant school characteristics year-specific trends, OSS rates are probably related to unobserved heterogeneity in test scores. While school FEs can sweep away stable characteristics of the school, one may expect that the shift away from zero-tolerance could lead schools to respond with changes in leadership, teacher behavior, student behavior, or a combination of reactions. It is unlikely that these changes would occur in parallel across schools, meaning they would not be captured by the inclusion of year FEs. Hence, the results of estimating Equation 1 will not necessarily hold a causal interpretation. These estimates are nonetheless meaningful descriptions which contribute to the existing literature investigating contributing factors to the observed trends in test scores. I provide further details as to why in Section 6.

Given the standardization that is performed on test scores, one must be interpret the magnitude of $\hat{\beta}_t$ carefully. The OSS rate variable is coded to vary from 0 to 1. Hence, we interpret $\hat{\beta}_t$ as the number of standard deviations a school's average test score is predicted to shift in response an increase in OSS rate from 0% to 100%.

I estimate two similar models at the student level. Let OSS_{it} be an indicator for whether student i receives any OSS in time period t. We consider specifications with both $t \in \{2008, \ldots, 2019\}$, and $t \in \{\text{Pre-2014}, \text{Post-2014}\}$. We estimate the following regression specification:

$$y_{it} = \alpha_t OSS_{it} + \xi_i + \lambda_s + \delta_t + \eta_{st} + \psi W_{it} + \nu_{it} \qquad (2)$$

where y_{it} is an individual student's test score in a given school year, ξ_i , λ_s, δ_t, and η_{st} are individual (i.e. student), school, school year, and school-by-year FEs, respectively, W_{it} is a vector of controls that include the same school-level controls as before but also student-level characteristics and demographics, and ν_{it} is the remaining unobserved heterogeneity in the model. Standard errors are clustered at the student level to account for the correlation in outcomes that come about as a result of each student appearing multiple times in the data.

Allowing α_t to vary with time enables an investigation of whether the relationship between individual OSS incidence and test scores has changed over the period of study, as well as specifically before versus after the 2014 Obama guidance. As with the school-level analysis, this flexibility allows for a more nuanced understanding of how the effects of OSS may have evolved alongside shifting policy and disciplinary norms. Specifically, heterogeneity across years of the estimates in the school- and student-level models taken together uncovers important temporal patterns in the OSS–achievement relationship that have not been documented in the existing literature. An alternative specification, which does not allow α_t to vary over time (i.e. I estimate α), is also presented in Section 6.

A similar identifying assumption to the one for Equation 1 must hold for Equation 2. As in the school-level analysis, this assumption is difficult to make. I do not claim that the inclusion of a student FE, which captures fixed student characteristics such as innate academic ability, propensity for disruptive behavior, and family background, is enough to sweep out any potential relationship between OSS incidence and unobserved heterogeneity in test scores. It is plausible at the student level that receiving an OSS in one year could affect the student's future behavior or attitudes toward suspension. Additionally, peer networks may be disrupted when a student is

suspended, introducing another potential source of endogeneity. Despite these concerns, the resulting estimates are still informative.

As a complement to Equation 2, we also estimate the following regression specification:

$$y_{it} = \theta_t OSS_{it} + \pi_t OSSRate_{sgt} + \xi_i + \lambda_s + \delta_t + \phi W_{it} + \varepsilon_{it}, \quad (3)$$

where y_{it}, W_{it}, ξ_i , λ_s, δ_t, and η_{st} are defined in the same way as in Equation 2, and ε_{it} is the remaining unobserved heterogeneity in the model. Note that Equation 3 does not include school-by-year FEs since including this would sweep away all variation in $OSSRate_{sgt}$. Standard errors are clustered at the student level to account for the correlation in outcomes that come about as a result of each student appearing multiple times in the data. As was the case for Equations 1 and 2, allowing for temporal heterogeneity in my coefficient estimates is crucial to my paper's contributions. However, an alternative specification in which estimates for θ and π are not allowed to vary with time can be useful to assess the importance of including student FEs and time-varying controls.

The school-level covariate $OSSRate_{sgt}$ in Equation 3 can be thought of as a proxy for overall disciplinary climate of a school, and its coefficient can be interpreted as a possible measure of peer effects. Its addition allows me to somewhat disentangle the direct consequences of a student's own suspension from the indirect consequences of being in a setting where suspensions are more (or less) prevalent. For instance, higher OSS rates might reflect a more punitive disciplinary culture, which could affect student morale, classroom disruptions, or even teacher behavior—all of which may in turn influence test performance. Conversely, they might also reflect underlying behavioral norms in a given school-grade cohort. By jointly estimating Equations 2 and 3, we can better assess the extent to which both individual and peer-level exposure to exclusionary discipline relate to student achievement.

Note that Equations 2 and 3 will be estimated using over 3 million observations. With such a large sample size, conventional frequentist significance testing becomes less informative, as even trivially small effects will typically achieve conventional levels of statistical significance due to the sheer volume of data. In large samples, the standard errors shrink, making virtually any coefficient statistically distinguishable from zero. To address this, I supplement classical hypothesis testing with a Bayesian model selection approach known as the Schwarz criterion (Schwarz, 1978). This approach penalizes model complexity and selects the model that is most likely a posteriori. Specifically, when comparing a restricted model with an unrestricted model, the critical value for evaluating whether to reject the restrictions is given by $r \cdot ln(N)$, where r is the number of restrictions and N is the sample size. For a single restriction, the critical value becomes $ln(N)$ and the test statistic is asymptotically distributed as the square of a t-distribution. In other words, hypothesis testing can be performed on individual coefficient estimates by comparing t statistics to a critical value of $\sqrt{ln(N)}$. For N = 3,000,000, $\sqrt{ln(N)} \approx 3.862$.

6 Results

6.A OSS Rates and School-Level Achievement

Table 1 presents estimates of β for Equation 1 under a series of nested specifications, with and without controls for school FEs, year FEs, and school compositional variables. It also presents, for comparison, estimates of Equation 1 without school and year FEs. The baseline specifications in Columns 1 and 7 show a large and highly significant negative association between OSS rates and average test scores in reading and math, with coefficients of –1.700 and –1.735, respectively (i.e. a change from 0% to 100% OSS rate predicts a 1.7 and 1.735 standard

Table 1. Mean test score and OSS rate

	Reading						**Math**					
	Base			EDS/Race Controls			Base			EDS/Race Controls		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
OSS Rate	-1.700***	-0.409***	-0.428***	-1.096***	-0.325***	-0.326***	-1.735***	-0.314***	-0.321***	-1.122***	-0.227***	-0.227***
	(-35.86)	(-8.01)	(-8.24)	(-27.85)	(-6.70)	(-6.68)	(-38.56)	(-6.41)	(-6.44)	(-30.50)	(-4.96)	(-4.91)
% Econ. Disadv.				-0.835***	-0.170***	-0.193***				-0.849***	-0.171***	-0.173***
				(-21.47)	(-5.50)	(-5.54)				(-16.37)	(-5.33)	(-4.93)
% Black				-0.112	-0.267**	-0.275**				-0.300**	-0.361***	-0.351**
				(-1.28)	(-2.88)	(-2.95)				(-2.95)	(-3.37)	(-3.25)
% Hispanic				-0.186*	-0.143	-0.107				-0.169	-0.123	-0.142
				(-2.18)	(-1.42)	(-1.04)				(-1.66)	(-1.02)	(-1.18)
% White				0.173*	0.321**	0.298**				-0.0120	0.181	0.196
				(1.98)	(3.25)	(2.95)				(-0.12)	(1.79)	(1.89)
Constant	0.177***	-0.0302***	-0.0272**	0.490***	-0.0282	-0.00630	0.160***	-0.0683***	-0.0670***	0.625***	0.0301	0.0230
	(12.46)	(-3.69)	(-3.27)	(5.78)	(-0.32)	(-0.07)	(10.39)	(-8.70)	(-8.38)	(6.25)	(0.31)	(0.24)
School FE		✓	✓		✓	✓		✓	✓		✓	✓
School Year FE			✓			✓			✓			✓
Observations	24949	24882	24882	24949	24882	24882	24931	24864	24864	24931	24864	24864

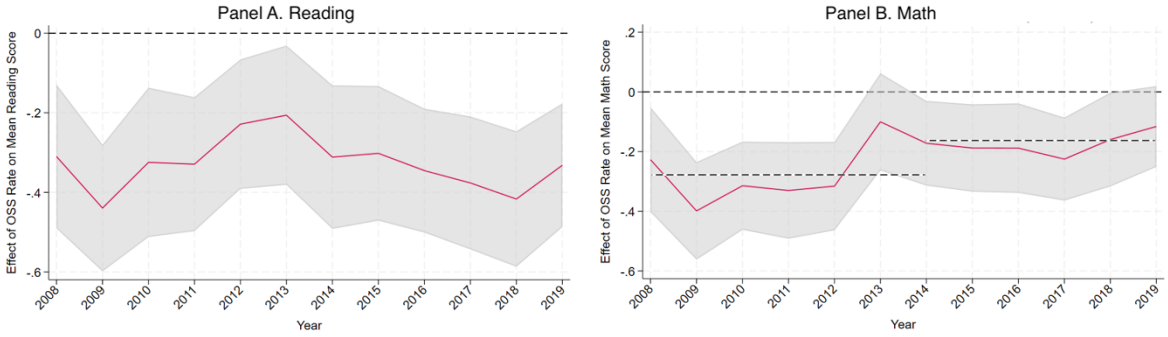
Notes. Results of estimating Equation 1 for reading and math score outcomes are included in Columns 3, 6, 9, and 12. Columns 1, 4, 7, and 10 are Equation 1 without any FEs, while Columns 2, 5, 8, and 11 include school FEs but do not control for year FEs. Base columns do not control for any other covariates, while EDS/Race Controls columns include school composition variables in the vector of controls W_{sgt}. The first row represents estimates of β under these specifications. t statistics using standard errors clustered by school are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

deviation drop in average test scores of a grade in a school in reading and math, respectively). These results reflect purely descriptive associations, without accounting for unobserved school-level heterogeneity or temporal variation. Once school FEs are included (Columns 2 and 8), the magnitude of the coefficients drops substantially—by more than 75%—suggesting that much of the initial correlation was driven by persistent differences across schools. Adding school year FEs in Columns 3 and 9 leads to only minimal further change.

Columns 4–6 and 10–12 introduce controls for time-varying school-level demographic composition (economic disadvantage and race shares). The inclusion of these controls leads to an additional reduction in the magnitude of the OSS coefficient, by roughly one-quarter in reading and one-third in math (comparing Columns 3 and 6, and 9 and 12, respectively). This pattern indicates that compositional factors account for additional variation in test scores that is correlated with OSS rates. Together, these results highlight the importance of accounting for both fixed school characteristics and demographic composition when estimating the relationship between OSS and academic performance.

Even after controlling for school and year FEs as well as school composition covariates, OSS rates remain significantly negatively associated with test scores. These findings run counter to what one might have expected from the descriptive trends shown earlier, where OSS rates and NAEP scores appeared to follow similar trajectories. Despite that parallel, Table 1 suggests that, within schools, higher OSS rates are still associated with lower academic performance. When I instead estimate the coefficient β_t in Equation 1 for $t = 2008, \ldots, 2019$, we gain a far more nuanced perspective. Figure 3 plots these estimates, allowing us to capture potential heterogeneity in the effect of OSS rates on test scores over time. Panel A shows the estimates for reading scores. Here, we observe relatively little variation across years. In contrast, Panel B

Figure 3. Mean test score and OSS rate by school year



Notes. Results $\hat{\beta}_t$ of estimating the coefficient β_t in Equation 1 for $t = 2008, \ldots, 2019$ for both reading and math scores. Point estimates and 99% confidence intervals are plotted. All estimates include school and school year FEs, with standard errors clustered at the school level. Panel B additionally plots the mean effect of pre- and post-2014. Panel A does not plot this since the pre- and post-2014 periods have essentially the same mean estimates. Appendix 2 provides estimates of Equation 1 for the pre- and post-2014 periods, where the mean effects that are currently plotted in Panel B can be seen.

reveals more pronounced heterogeneity in the math estimates. In particular, we see a noticeable upward shift in 2013, one year prior to the 2014 federal guidance, but several years after state-level reforms began in North Carolina and OSS rates had already declined substantially. The effect sizes become far less negative during this period, and in 2013, 2018, and 2019, the coefficients are not statistically distinguishable from zero at the 1% level. Moreover, the average effect before and after 2014 differs significantly at the 0.1% level (see Appendix 2).

Given that OSS rates had already declined substantially by 2013 (the year in which we observe the most notable shift in the estimated effect) from their 2010 peak, the results raise the possibility that schools which reduced suspension rates earlier may have been more likely to experience subsequent improvements in test scores. While this pattern is suggestive, formally testing this hypothesis lies beyond the scope of the present analysis.

While I am cautious not to interpret these results causally—see discussion in Section 5—they nevertheless contribute a novel insight to the existing literature. Prior work has largely drawn conclusions from identification of a single causal effect of suspension on academic performance. These findings complicate that narrative by showing that the estimated effects of

OSS rates on achievement are not stable over time. Panel B in particular suggests that in high-suspension environments like 2008, the marginal impact of an additional suspension may differ substantially from the impact in lower-suspension environments in the years closer to 2019. These findings hence align with the spirit of a MTE framework. Even though the effect is never positive, by the later years of my sample, it is closer to zero.

Appendix 3 provides evidence that these effects are also heterogeneous across school types. In particular, schools that qualify for Title I funding, small schools, and majority Black schools all exhibit more negative estimates for math scores prior to 2014 compared to after, with differences significant at the 1% level. Interestingly, large schools show negative effects after 2014, while pre-2014 estimates are not significantly different from zero. For majority Black and Title I schools, post-2014 estimates are not statistically distinguishable from zero. I keep the discussion of these results brief, as several zero-effect estimates may be influenced by limited statistical power when analyzing smaller subsamples. For instance, only about one-sixth of the sample consists of majority Black schools. No comparable heterogeneity is observed in the reading results.

Together, these results yield two key insights. First, they suggest that the large statewide reductions in OSS rates during this period do not appear to have worsened academic performance. Second, they highlight that the association between OSS and achievement may be context-specific and evolve over time. This stands in contrast to much of the literature, which treats suspension effects as time-invariant. To probe this issue more directly—and to explore whether the underlying effect of suspension on individuals has changed—we now turn to student-level analysis in Section 6.B.

Table 2. Individual test score and OSS incident, OSS rate

	Reading								**Math**							
	Base				School EDS/Race % Controls				Base				School EDS/Race % Controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
OSS	-0.738★	-0.549★	-0.0666★	-0.0628★	-0.573★	-0.550★	-0.0690★	-0.0629★	-0.755★	-0.551★	-0.0859★	-0.0814★	-0.582★	-0.553★	-0.0881★	-0.0814★
	(-262.81)	(-192.01)	(-46.54)	(-43.27)	(-210.31)	(-194.99)	(-48.10)	(-43.34)	(-302.90)	(-213.75)	(-64.59)	(-59.81)	(-239.89)	(-216.77)	(-65.54)	(-59.82)
OSS Rate		-1.869★		-0.224★		-0.293★		-0.219★		-2.018★		-0.243★		-0.392★		-0.243★
		(-182.82)		(-28.80)		(-30.57)		(-27.97)		(-201.89)		(-32.25)		(-42.94)		(-32.11)
% Econ. Disadv.					-1.119★	-1.069★	-0.0403★	-0.00728					-1.173★	-1.142★	0.0394★	0.0483★
					(-192.89)	(-192.42)	(-7.04)	(-1.43)					(-202.27)	(-198.37)	(7.94)	(9.71)
% Black					-0.305★	-0.274★	0.0389	-0.00437					-0.511★	-0.467★	-0.00833	0.0125
					(-21.57)	(-20.09)	(1.53)	(-0.25)					(-37.62)	(-34.31)	(-0.48)	(0.73)
% Hispanic					-0.274★	-0.292★	0.188★	0.129★					-0.188★	-0.200★	0.157★	0.144★
					(-17.31)	(-19.27)	(6.93)	(6.57)					(-12.46)	(-13.30)	(8.23)	(7.57)
% White					-0.0575★	-0.0559★	0.0953★	0.0151					-0.229★	-0.238★	0.0590***	0.0499**
					(-4.38)	(-4.44)	(3.93)	(0.88)					(-17.93)	(-18.70)	(3.53)	(2.98)
Constant	0.0684★	0.200★	0.0176★	0.0346★	0.797★	0.786★	-0.0479*	0.0124	0.0680★	0.209★	0.0169★	0.0352★	0.956★	0.962★	-0.0557***	-0.0407**
	(55.75)	(139.10)	(162.45)	(57.80)	(65.57)	(67.40)	(-2.22)	(0.82)	(54.25)	(141.93)	(165.85)	(60.86)	(79.89)	(80.76)	(-3.75)	(-2.74)
Student FE			✓	✓			✓	✓			✓	✓			✓	✓
School FE			✓	✓			✓	✓			✓	✓			✓	✓
School Year FE			✓	✓			✓	✓			✓	✓			✓	✓
School × Year FE			✓								✓					
Observations	3058431	3058400	3057517	3058380	3058431	3058400	3058399	3058380	3048263	3048263	3047378	3048242	3048263	3048263	3048242	3048242
Schwarz Critical Value	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864	3.864

Notes. Estimates for Equations 2 and 3 with and without FEs and controls for school composition. First row of odd numbered columns displays estimates for α from Equation 2. First row of even numbered columns displays estimates for θ from Equation 3. Second row displays estimates for π from Equation 3. t statistics using standard errors clustered at the student level are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ★ $t > \sqrt{\log(N)}$

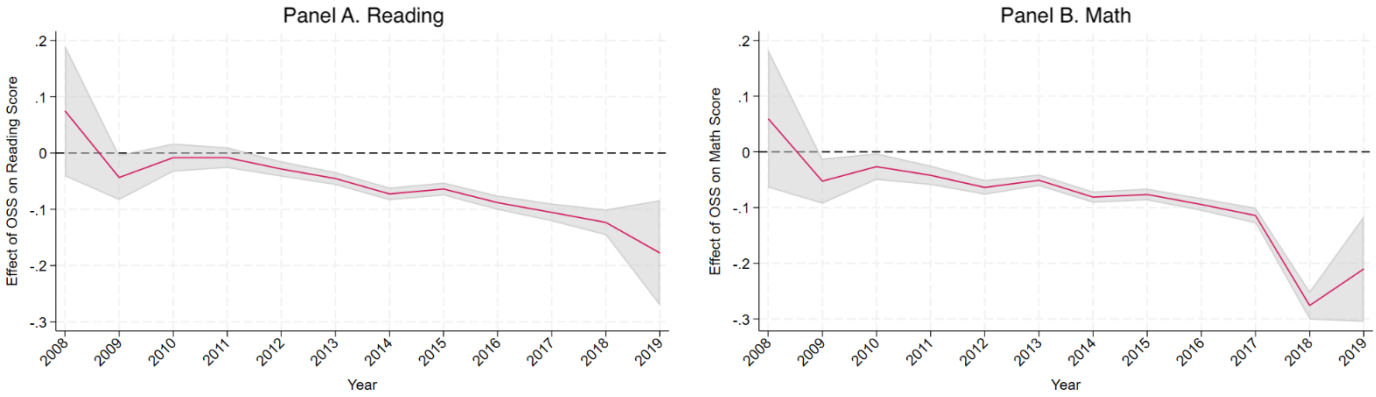
6.B OSS Incidence and Student-Level Achievement

Just as in Table 1, Table 2 presents estimates of α for Equation 2 and θ and π for Equation 3 under a series of specifications, with and without controls for student, school, year, and school-by-year FEs, as well as school composition variables. As in the school-level analysis, the coefficients in Columns 1 and 9 represent a purely descriptive specification without any FEs, yielding large negative effect sizes—suggesting a strong negative association between OSS and academic outcomes. We can think of Columns 2 and 10 as including a control for potential peer effects, while remaining purely descriptive. However, once student, school, year (and school-by-year when applicable) FEs are included (Columns 3, 4, 11, and 12), these estimates shrink substantially, with the OSS coefficients dropping below 0.07 standard deviations in magnitude for reading and 0.09 for math.

Adding controls for school composition in Columns 5–8 and 13–16 has relatively modest effects on coefficient size, especially when compared to the school-level analysis. This suggests that the inclusion of student FEs accounts for a substantial portion of the unobserved heterogeneity that might otherwise bias the estimates. Even so, the estimates indicate that receiving an OSS is still associated with a small but statistically significant decrease in test scores, conditional on student, school, and time-invariant characteristics. This motivates further exploration into whether these effects vary meaningfully over time or by student subgroup.

When I instead estimate θ_t from Equation 3 for each year from 2008 to 2019, a more nuanced and revealing picture emerges, as shown in Figure 4. In sharp contrast with the results at the school level, in the student-level models for both reading (Panel A) and math (Panel B), the estimated effect of OSS on test scores becomes more negative over time. Notably, the estimated

Figure 4. Individual test score and OSS by school year



Notes. Results $\hat{\theta}_t$ of estimating the coefficient θ_t in Equation 3 for $t = 2008, \ldots, 2019$ for both reading and math scores. Point estimates and confidence intervals based on Schwarz criterion critical values are plotted. All estimates include student, school, and school year FEs, as well as controls for time-varying school compositional variables, with standard errors clustered at the student level.

effect of receiving an OSS in 2008 is not significantly different from zero for either subject, with a null result for reading throughout 2009–2011 as well. These results are broadly consistent across student demographics, suggesting relatively limited heterogeneity in the extent to which different groups are affected. The only exceptions, shown in Appendix 4 and Appendix 5, are for White and Hispanic students, for whom the math estimates are not significantly different from zero until 2012. In addition, for Hispanic students, the reading estimates remain null until 2014.

These results are in stark contrast to the pattern observed in the school-level results from Figure 3. This divergence between the school- and student-level patterns underscores the importance of moving to individual-level analysis. Whereas the school-level estimates suggested that the effects of OSS on academic performance may have weakened over time (particularly for math), the student-level estimates suggest the opposite: the academic consequences of receiving an OSS appear to have grown more severe.

One possible interpretation is that as schools reduced their overall use of OSS, the students who continued to receive suspensions may have been those involved in more serious or disruptive incidents. In earlier years, when OSS rates were higher (e.g. exceeding 15% among

middle schoolers in 2008), receiving a suspension may have carried less stigma and been less indicative of severe behavioral problems. As such, being suspended may have had less impact on classroom dynamics or peer/teacher perception, and therefore on academic outcomes. By 2019, however, with middle school OSS rates nearing 10%, being suspended may have signaled more serious misconduct. This could have resulted in students being treated differently by teachers, peers, or administrators, thereby exacerbating the academic consequences of the suspension itself. Another possibility is that, as OSS rates have decreased, schools' use of suspension has become more targeted, increasingly reserved for students involved in the most serious forms of misbehavior. In this context, those who are still suspended may differ systematically from those who were suspended in earlier years along dimensions that could independently contribute to worse academic outcomes, even after controlling for time-invariant student characteristics.

The temporal heterogeneity observed in Figure 4 offers a crucial empirical insight: the OSS–achievement relationship is not stable over time, calling into question the immediate relevance of past findings that focus on a specific historical moment, particularly one that may not reflect the disciplinary landscape students face today.

Importantly, these results cannot be explained by time-invariant student characteristics, as these are accounted for by the inclusion of student FEs. Nor can they be dismissed as school-level shifts, as the school FEs absorb any persistent differences across schools. What remains, then, appears to be a growing relationship between receiving an OSS and experiencing academic decline—one that is tightly linked to the evolving role and meaning of OSS itself. This pattern would be invisible in purely school-level analyses. As OSS becomes less frequent, the suspended population shrinks, and their negative outcomes are diluted in school-level averages. Moreover, the smaller number of students who are being suspended are likely to be, on average, far more

disruptive than students being suspended earlier in our sample period. Hence, their removal from schools could improve peer achievements. The more muted school-level estimates in recent years, as discussed in Section 6.A could reflect a combination of these effects.

Before assigning a causal interpretation to these estimates, it is worth testing a key assumption that is necessary for that interpretation. It remains possible that the same behavioral disruptions that trigger an OSS also independently drive poor academic outcomes, and that this relationship is growing stronger over time. Appendix 6 shows coefficient estimates for a regression of student test scores in year t on indicators for OSS receipt in year $t-2, t-1, t, t+1$, and $t+2$, for all t, controlling for student, school, and year FEs along with time-varying school composition variables. I find that OSS in future years is significantly predictive of lower present test scores, even after controlling for past and present OSS, suggesting the presence of negative selection. Students with worse academic outcomes are more likely to receive OSS in both the past and future, even after controlling for time-invariant student and school characteristics and common time trends. This implies that the coefficient of OSS is likely to be biased downward, i.e. it overstates the magnitude of the true causal effect of OSS on test scores.

This helps contextualize the magnitudes reported in Table 2. For example, in Columns 3 and 11, which account for student, school, year, and school-by-year FEs, I respectively estimate the effect of OSS on reading and math scores as –0.0666 and –0.0859 standard deviations. As Appendix 6 suggests, these estimates should not be seen as causal, and are instead likely to overstate the true causal effects due to negative selection. This allows us to plausibly interpret these estimates as upper bounds on the magnitude of the impact of OSS on academic achievement. This bounding exercise is especially important in light of the substantially larger estimates found in Columns 1 and 9, respectively –0.738 and –0.755, which omit FEs and reflect

purely descriptive associations. By offering credible bounds on the size of the effect, my analysis provides a significant contribution by clarifying that the true causal impact of OSS is likely far more limited than previously suggested.

6.C Discussion

Taken together, the results from Sections 6.A and 6.B point to two central findings. First, the academic consequences of OSS are not stable over time: at the school level, the relationship between OSS rates and achievement appears to weaken, while at the student level, the relationship becomes more negative. Second, these trends are not driven by persistent differences across schools or students, but instead reflect evolving dynamics in how OSS is used and perceived.

The effect of OSS on academic outcomes is shaped by context. As OSS rates decline, the meaning of receiving a suspension may shift dramatically. These patterns are consistent with the intuition behind a MTE framework, in which the impact of a treatment depends on the characteristics of the marginal recipients and the institutional environment in which it is applied.

While the estimates presented here should not be interpreted as strictly causal, they carry important implications for future research. The results establish credible upper bounds on the magnitude of the overall causal effect of OSS. These bounds are substantially smaller than those reported in much of the observational literature. Moreover, rather than searching for a single, time-invariant estimate of the suspension–achievement relationship, researchers should explore how the academic consequences of OSS evolve as disciplinary practices change. Continued work will be essential to understanding the conditions under which suspension policies help or hinder student learning.

7 Conclusion

This paper is motivated by a growing concern over the sustained decline in student achievement in the United States, as measured by standardized assessments such as the National Assessment of Educational Progress (NAEP). Using rich administrative data from the North Carolina Education Research Data Center (NCERDC), I investigate whether changes in exclusionary discipline practices, specifically out-of-school suspensions (OSS), can help explain these troubling academic trends.

My analysis contributes to two related literatures. First, I add to research examining potential explanations for broad patterns in test scores by highlighting the role of a factor that is both manipulable by policy and relatively low-cost to address. Second, I contribute to the small but growing body of causal work on the academic effects of OSS. I provide relatively tight upper bounds on the magnitude of the overall causal effect of OSS at the student-level. Furthermore, while prior studies typically use data from the early-2010s or earlier, when zero-tolerance policies were still widely in place, I examine a later period marked by substantial disciplinary reform away from exclusionary practices accompanied by a steep decline in OSS rates. Drawing inspiration from the marginal treatment effect (MTE) framework, I explore whether the academic impact of OSS varies depending on how likely students are to be suspended. While I do not formally estimate MTEs, the analysis is motivated by the underlying intuition of the framework: as schools suspend fewer students overall, the marginal student who receives an OSS may differ systematically from those suspended in the past, potentially changing the nature of the treatment effect. I show that the estimated relationship between OSS and achievement is not stable over time. In fact, while the school-level effects of OSS appear to weaken in more recent years, the

student-level effects become increasingly negative. These findings suggest that as schools suspend fewer students overall, OSS becomes an extreme negative signal, leading to stigma and differential treatment, ultimately worsening outcomes. Schools' use of suspensions may have also become more selective in recent years, meaning those who are still suspended differ in meaningful ways from students who would have been suspended in earlier years—differences that may contribute to the increasingly negative estimates in the student-level analysis. Through the same mechanism, negative school-level estimates shrink as suspensions now target a smaller group of students who are more likely to be disruptive.

Looking ahead, several directions for future work emerge from this analysis. First, a fuller implementation of a MTE framework would allow for explicit estimation of how the effects of OSS vary across different margins of disciplinary behavior. Second, examining racial disparities in both the assignment and consequences of OSS could help shed light on the equity implications of current policies. Finally, future work could explore heterogeneity in OSS effects along two additional dimensions: (1) the severity of the punishment, measured by suspension length, and (2) the type of behavior that led to the suspension. Together, these extensions would provide a more granular understanding of how exclusionary discipline shapes academic outcomes—and for whom.

8 References

American Psychological Association Zero Tolerance Task Force. (2008). Are zero tolerance policies effective in the schools?: An evidentiary review and recommendations. *The American Psychologist*, *63*(9), 852–862. https://doi.org/10.1037/0003-066X.63.9.852

Anderson, K. P., Ritter, G. W., & Zamarro, G. (2017). *Understanding a Vicious Cycle: Do Out-of-School Suspensions Impact Student Test Scores?* (Education Reform Faculty and Graduate Students Publications). https://scholarworks.uark.edu/edrepub/11

Arcia, E. (2006). Achievement and Enrollment Status of Suspended Students: Outcomes in a Large, Multicultural School District. *Education and Urban Society*, *38*(3), 359–369. https://doi.org/10.1177/0013124506286947

Baron, E. J. (2022). School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin. *American Economic Journal: Economic Policy*, *14*(1), 1–39. https://doi.org/10.1257/pol.20200226

Campbell, F. A., & Ramey, C. T. (1994). Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families. *Child Development*, *65*(2), 684–698. JSTOR. https://doi.org/10.2307/1131410

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., & Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educational Research Review*, *14*, 33–46. https://doi.org/10.1016/j.edurev.2015.01.002

Cerrone, K. M. (1999). The Gun-Free Schools Act of 1994: Zero Tolerance Takes Aim at Procedural Due Process. *Pace Law Review*, *20*(1), 131.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Cook, P. J., & Kang, S. (2016). Birthdays, Schooling, and Crime: Regression-Discontinuity Analysis of School Performance, Delinquency, Dropout, and Crime Initiation. *American*

Economic Journal: Applied Economics, *8*(1), 33–57. https://doi.org/10.1257/app.20140323

Council of State Governments Justice Center. (2017). *Realizing the Full Vision of School Discipline Reform: A Framework for Statewide Change*. The Council of State Governments Justice Center. https://csgjusticecenter.org/publications/realizing-the-full-vision-of-school-discipline-reform-a-framework-for-statewide-change/

Currie, J., & Thomas, D. (1993). Does Head Start Make a Difference? *National Bureau of Economic Research Working Paper Series*, *No. 4406*. https://doi.org/10.3386/w4406

Dee, T. S., & Jacob, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. https://doi.org/10.1002/pam.20586

Dogan, U. (2015). Student engagement, academic self-efficacy, and academic motivation as predictors of academic performance. *The Anthropologist*, *20*(3), 553–561.

Donnelly, J. E., Hillman, C. H., Castelli, D., Etnier, J. L., Lee, S., Tomporowski, P., Lambourne, K., & Szabo-Reed, A. N. (2016). Physical Activity, Fitness, Cognitive Function, and Academic Achievement in Children: A Systematic Review. *Medicine and Science in Sports and Exercise*, *48*(6), 1197–1222. https://doi.org/10.1249/MSS.0000000000000901

Doty, E., Kane, T. J., Patterson, T., & Staiger, D. O. (2022). What Do Changes in State Test Scores Imply for Later Life Outcomes? *National Bureau of Economic Research Working Paper Series*, *No. 30701*. https://doi.org/10.3386/w30701

Duckworth, A. L., Taxer, J. L., Eskreis-Winkler, L., Galla, B. M., & Gross, J. J. (2019). Self-Control and Academic Achievement. In *Annual Review of Psychology* (Vol. 70, Issue

Volume 70, 2019, pp. 373–399). Annual Reviews. https://doi.org/10.1146/annurev-psych-010418-103230

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class Size and Student Achievement. *Psychological Science in the Public Interest*, *2*(1), 1–30. https://doi.org/10.1111/1529-1006.003

Ewing, C. P. (2000). *Sensible Zero Tolerance Protects Students*. Harvard Education Letter: Research Online. https://files.eric.ed.gov/fulltext/ED456938.pdf

Finn, J. D., & Achilles, C. M. (1990). Answers and Questions about Class Size: A Statewide Experiment. *American Educational Research Journal*, *27*(3), 557–577. JSTOR. https://doi.org/10.2307/1162936

Gajda, A., Karwowski, M., & Beghetto, R. A. (2017). Creativity and Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, *109*(2), 269–299. https://doi.org/10.1037/edu0000133

Garces, E., Thomas, D., & Currie, J. (2002). Longer-Term Effects of Head Start. *American Economic Review*, *92*(4), 999–1012. https://doi.org/10.1257/00028280260344560

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). Does Head Start Do Any Lasting Good? *National Bureau of Economic Research Working Paper Series*, *No. 17452*. https://doi.org/10.3386/w17452

Grissmer, D. W., Flanagan, A., Kawata, J. H., & Williamson, S. (2000). *Improving Student Achievement: What State NAEP Test Scores Tell Us*. RAND Corporation. https://doi.org/10.7249/MR924

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student Achievement and the Changing American Family*. RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR488.html

Hanushek, E. A. (1999). Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, *21*(2), 143–163. https://doi.org/10.3102/01623737021002143

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051–1078.

Hanushek, E. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the Estimated Effects of School Resources. *The Review of Economics and Statistics*, *78*(4), 611–627. JSTOR. https://doi.org/10.2307/2109949

Hanushek, E. A., & Woessmann, L. (2017). School Resources and Student Achievement: A Review of Cross-Country Economic Research. In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson* (pp. 149–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-43473-5_8

Havard, B., Nguyen, G.-N., & Otto, B. (2018). The impact of technology use and teacher professional development on U.S. national assessment of educational progress (NAEP) mathematics achievement. *Education and Information Technologies*, *23*(5), 1897–1918. https://doi.org/10.1007/s10639-018-9696-4

Hoffman, S. (2014). Zero Benefit: Estimating the Effect of Zero Tolerance Discipline Polices on Racial Disparities in School Discipline. *Educational Policy*, *28*(1), 69–95. https://doi.org/10.1177/0895904812453999

Hwang, N., & Domina, T. (2021). Peer Disruption and Learning: Links between Suspensions and the Educational Achievement of Non-Suspended Students. *Education Finance and Policy*, *16*(3), 443–463. https://doi.org/10.1162/edfp_a_00308

Jackson, C. K., Johnson, R., & Persico, C. (2014). The Effect of School Finance Reforms on the Distribution of Spending, Academic Achievement, and Adult Outcomes. *National Bureau of Economic Research Working Paper Series*, *No. 20118*. https://doi.org/10.3386/w20118

Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do School Spending Cuts Matter? Evidence from the Great Recession. *American Economic Journal: Economic Policy*, *13*(2), 304–335. https://doi.org/10.1257/pol.20180674

Lacoe, J., & Steinberg, M. P. (2019). Do Suspensions Affect Student Outcomes? *Educational Evaluation and Policy Analysis*, *41*(1), 34–62. https://doi.org/10.3102/0162373718794897

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, *10*(2), 1–26. https://doi.org/10.1257/app.20160567

Lee, V. E., Brooks-Gunn, J., Schnur, E., & Liaw, F.-R. (1990). Are Head Start Effects Sustained? A Longitudinal Follow-Up Comparison of Disadvantaged Children Attending Head Start, No Preschool, and Other Preschool Programs. *Child Development*, *61*(2), 495–507. JSTOR. https://doi.org/10.2307/1131110

Leung-Gagné, M., McCombs, J., Scott, C., & Losen, D. J. (2022). *Pushed Out: Trends and Disparities in Out-of-School Suspension*. Learning Policy Institute. https://doi.org/10.54300/235.277

Lochner, L. (2020). Chapter 9—Education and crime. In S. Bradley & C. Green (Eds.), *The Economics of Education (Second Edition)* (pp. 109–117). Academic Press. https://doi.org/10.1016/B978-0-12-815391-8.00009-4

Ludwig, J., & Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*. *The Quarterly Journal of Economics*, *122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE Program: A Pilot Program in Targeted Pupil-Teacher Reduction in Wisconsin. *Educational Evaluation and Policy Analysis*, *21*(2), 165–177. https://doi.org/10.3102/01623737021002165

Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades. *The Future of Children*, *5*(2), 113–127. JSTOR. https://doi.org/10.2307/1602360

Noltemeyer, A. L., Ward ,Rose Marie, & and Mcloughlin, C. (2015). Relationship Between School Suspension and Student Outcomes: A Meta-Analysis. *School Psychology Review*, *44*(2), 224–240. https://doi.org/10.17105/spr-14-0008.1

North Carolina Department of Public Instruction. (2021, February). *North Carolina Discipline Data Reporting Procedures*. https://www.dpi.nc.gov/discipline-data-reporting-procedures/open

North Carolina Department of Public Instruction. (2025). *End-of-Grade (EOG)*.

Robson, D. A., Allen, M. S., & Howard, S. J. (2020). Self-regulation in childhood as a predictor of future outcomes: A meta-analytic review. *Psychol Bull*, *146*(4), 324–354. PubMed. https://doi.org/10.1037/bul0000227

Schanzenbach, D. W. (2006). What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy*, *9*, 205–228. JSTOR.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. JSTOR.

Skiba, R., & Peterson, R. (1999). The Dark Side of Zero Tolerance: Can Punishment Lead to Safe Schools? *The Phi Delta Kappan*, *80*(5), 372–382. JSTOR.

Sorensen, L. C., Bushway, S. D., & Gifford, E. J. (2022). Getting Tough? The Effects of Discretionary Principal Discipline on Student Outcomes. *Education Finance and Policy*, *17*(2), 255–284. https://doi.org/10.1162/edfp_a_00341

Steinberg, M. P., & Lacoe, J. (2018). Reforming School Discipline. *American Journal of Education*, *125*(1), 29–77. JSTOR.

U.S. Department of Education. (2025, January 14). *Every Student Succeeds Act (ESSA)*. U.S. Department of Education. https://www.ed.gov/laws-and-policy/laws-preschool-grade-12-education/every-student-succeeds-act-essa

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, & National Assessment of Educational Progress (NAEP). (2024a). *Math Assessment (2024, 2022, 2019, 2017, 2015, 2013, 2011, 2009, 2007, 2005, 2003, 2000, 1996, and 1992)* [Dataset].

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, & National Assessment of Educational Progress (NAEP). (2024b). *Reading*

Assessment (2024, 2022, 2019, 2017, 2015, 2013, 2011, 2009, 2007, 2005, 2003, 2000, 1996, and 1992) [Dataset].

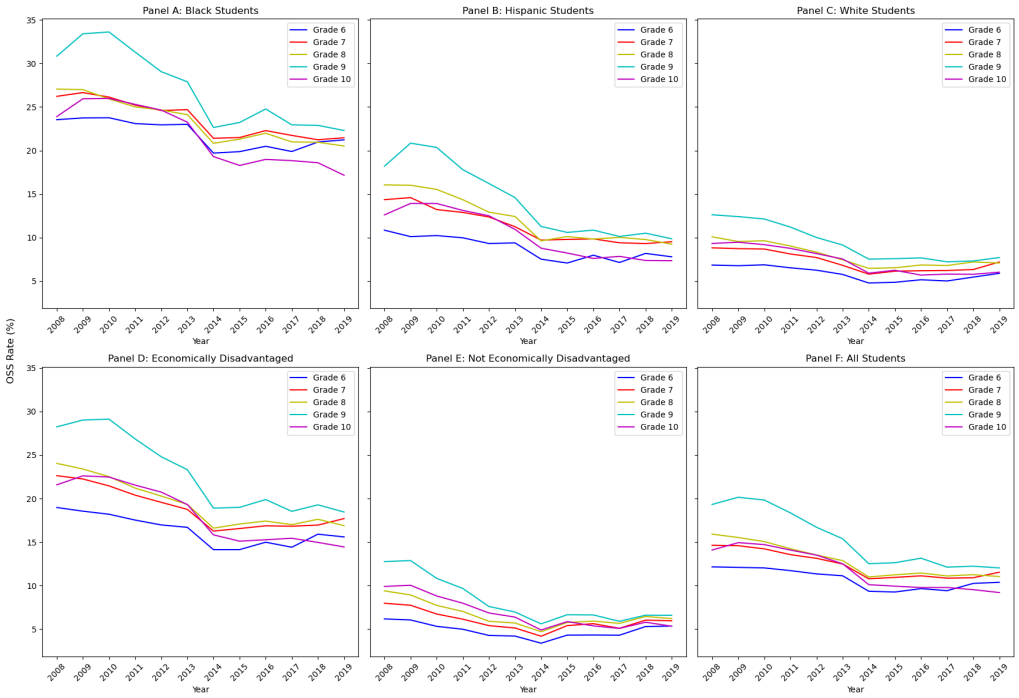
U.S. Department of Justice & U.S. Department of Education. (2014, January 8). *Departments of Justice and Education Issue School Discipline Guidance to Promote Safe, Inclusive Schools*. Office of Public Affairs. https://www.justice.gov/archives/opa/pr/departments-justice-and-education-issue-school-discipline-guidance-promote-safe-inclusive

Yaluma, C. B., Little, A. P., & Leonard, M. B. (2022). Estimating the Impact of Expulsions, Suspensions, and Arrests on Average School Proficiency Rates in Ohio Using Fixed Effects. *Educational Policy*, *36*(7), 1731–1758. https://doi.org/10.1177/0895904821999838

Zhai, F., Raver, C. C., & Jones, S. M. (2012). Academic performance of subsequent schools and impacts of early interventions: Evidence from a randomized controlled trial in Head Start settings. *Children and Youth Services Review*, *34*(5), 946–954. https://doi.org/10.1016/j.childyouth.2012.01.026

9 Appendix

Appendix 1. OSS Rate by Grade by Demographic Group (2008–2019, shared y-axis)



Notes. Panel A, B, and C plot OSS rate by grade (for grades 6–10) for Black, Hispanic, and White students, respectively. Panel D and E plot OSS rate by grade for economically disadvantaged students and non-disadvantaged students, respectively. Panel F shows overall OSS rate over time by grade. The vertical axes of these six panels are shared, which demonstrates the differences in OSS rate levels across demographic groups. The year on the horizontal axes denotes the year of the spring semester for each school year (e.g. 2013 denotes the 2012-13 school year).

Appendix 2. Mean math score and OSS rate by pre- vs. post-2014

	Base (1)	EDS/Race Controls (2)
OSS Rate	-0.264***	-0.176***
	(-5.18)	(-3.62)
% Econ. Disadv.		-0.167***
		(-4.76)
% Black		-0.352**
		(-3.26)
% Hispanic		-0.134
		(-1.11)
% White		0.196
		(1.90)
Pre-2014 × OSS Rate	-0.114***	-0.102***
	(-3.89)	(-3.49)
Constant	-0.0665***	0.0195
	(-8.33)	(0.20)
School FE	✓	✓
School Year FE	✓	✓
Observations	24864	24864

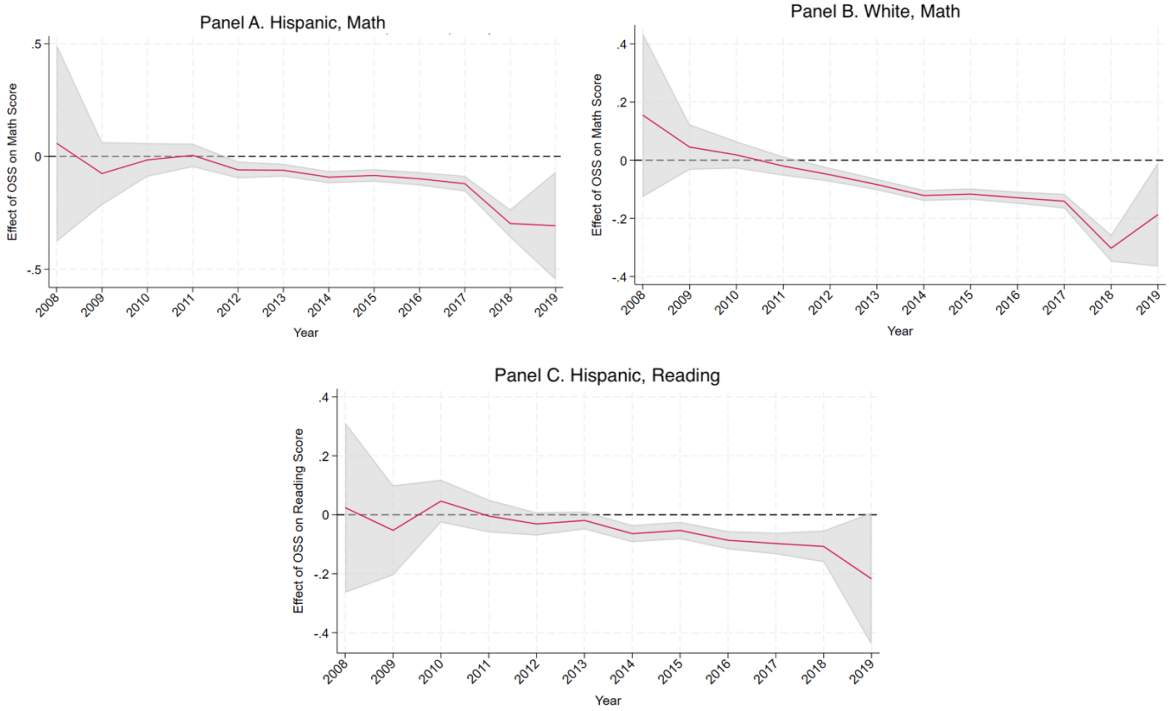
Notes. Results of estimating Equation 1, where mean math score is the outcome, with an interaction term Pre-2014 × OSS Rate. This is the same as estimating β_t for $t \in \{\text{Pre-2014}, \text{Post-2014}\}$, where $\hat{\beta}_{\text{Post-2014}}$ is equal to the estimate reported in the first row, and $\hat{\beta}_{\text{Pre-2014}}$ is the sum of the estimate in the first row and the estimate of the coefficient on the interaction term. t statistics using standard errors clustered by school are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix 3. Mean math score and OSS rate by pre- vs. post-2014 period, by school type

	All (1)	Title I (2)	Not Title I (3)	Rural (4)	Town (5)	City (6)	Small (7)	Mid-Sized (8)	Large (9)	Maj. Black (10)
OSS Rate	-0.176***	-0.230***	-0.379	-0.256***	-0.0669	-0.231**	-0.0310	-0.500***	-0.779***	-0.00767
	(-3.62)	(-4.51)	(-1.77)	(-3.85)	(-0.97)	(-3.00)	(-0.49)	(-6.79)	(-7.46)	(-0.10)
% Econ. Disadv.	-0.167***	-0.166***	-0.296***	-0.164***	-0.184**	-0.184**	-0.106*	-0.159**	-0.335***	0.0105
	(-4.76)	(-4.19)	(-3.95)	(-3.98)	(-3.18)	(-3.26)	(-2.15)	(-3.05)	(-5.96)	(0.20)
% Black	-0.352**	-0.245*	-0.654*	-0.420**	0.0127	-0.200	-0.299**	-0.0260	-0.948***	-0.492**
	(-3.26)	(-2.20)	(-2.52)	(-3.24)	(0.10)	(-1.01)	(-2.76)	(-0.11)	(-5.20)	(-3.09)
% Hispanic	-0.134	-0.0177	-0.339	-0.314*	0.296	0.0259	-0.0263	0.249	-0.738***	-0.306
	(-1.11)	(-0.13)	(-1.41)	(-2.26)	(1.63)	(0.12)	(-0.22)	(0.94)	(-4.02)	(-1.78)
% White	0.196	0.227*	0.00348	0.181	0.360**	0.325	0.0626	0.754***	-0.188	0.0263
	(1.90)	(2.06)	(0.02)	(1.41)	(2.61)	(1.78)	(0.56)	(3.45)	(-1.03)	(0.15)
Pre-2014 × OSS Rate	-0.102***	-0.113***	0.129	-0.0692	-0.118*	-0.105*	-0.111**	0.204*	0.304**	-0.163**
	(-3.49)	(-3.83)	(0.54)	(-1.32)	(-1.99)	(-2.08)	(-2.69)	(2.26)	(3.01)	(-3.08)
Constant	0.0195	-0.113	0.467*	0.0763	-0.283*	-0.0554	-0.389***	-0.382	0.700***	-0.153
	(0.20)	(-1.09)	(2.16)	(0.64)	(-2.14)	(-0.32)	(-3.58)	(-1.85)	(4.25)	(-1.01)
School FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	24864	18109	6719	17026	6938	9238	3921	9888	11038	4602

Notes. Results of estimating Equation 1 separately for different school types, where mean math score is the outcome, with an interaction term Pre-2014 × OSS Rate. This is the same as estimating β_t for $t \in \{\text{Pre-2014, Post-2014}\}$, where $\hat{\beta}_{\text{Post-2014}}$ is equal to the estimate reported in the first row, and $\hat{\beta}_{\text{Pre-2014}}$ is the sum of the estimate in the first row and the estimate of the coefficient on the interaction term. Small and Large schools denote schools in the bottom- and top-25% of schools when ranked by school population. Rural combines the suburb and rural classifications, since their definition was changed sometime during my sample period. NCERDC uses classification criteria defined by the Census Bureau. t statistics using standard errors clustered by school are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix 4. Test score and OSS for Hispanic and White students



Notes. Results $\hat{\theta}_t$ of estimating the coefficient θ_t in Equation 3 for $t = 2008, \ldots, 2019$. Panel A and B have math score as the outcome and Panel C has reading score as the outcome. Panel A and C are on the subset of Hispanic students; Panel B is on the subset of White students. Point estimates and 99% confidence intervals are plotted. All estimates include student, school, and school year FEs, with standard errors clustered at the student level.

Appendix 5. Test score and OSS by student demographic subgroup and year

	All		Black		Hispanic		White		Econ. Disadv.	
	Reading (1)	Math (2)	Reading (3)	Math (4)	Reading (5)	Math (6)	Reading (7)	Math (8)	Reading (9)	Math (10)
OSS	-0.174★	-0.209★	-0.153★	-0.177★	-0.217***	-0.307★	-0.197★	-0.187★	-0.191★	-0.211★
	(-7.19)	(-8.51)	(-4.63)	(-5.09)	(-3.49)	(-4.62)	(-3.84)	(-3.95)	(-6.56)	(-7.25)
OSS Rate	-0.225★	-0.251★	-0.253★	-0.261★	-0.307★	-0.303★	-0.166★	-0.264★	-0.269★	-0.261★
	(-28.80)	(-33.08)	(-21.53)	(-22.65)	(-15.59)	(-15.37)	(-12.12)	(-20.26)	(-27.72)	(-27.86)
% Econ. Disadv.	-0.0101*	0.0445★	0.00414	0.0331★	0.000156	0.0308*	0.0308★	0.0620★	0.0199**	0.0389★
	(-1.98)	(8.93)	(0.46)	(3.72)	(0.01)	(2.41)	(3.90)	(8.24)	(2.65)	(5.31)
% Black	-0.00174	0.0137	-0.0289	-0.0352	0.0123	0.00808	-0.0373	0.00547	-0.0137	-0.0188
	(-0.10)	(0.79)	(-0.91)	(-1.14)	(0.26)	(0.17)	(-1.32)	(0.20)	(-0.56)	(-0.79)
% Hispanic	0.131★	0.147★	0.171★	0.164★	0.108*	0.0917	0.0656*	0.0698*	0.109★	0.108★
	(6.70)	(7.69)	(4.72)	(4.63)	(2.20)	(1.86)	(2.15)	(2.37)	(4.05)	(4.17)
% White	0.0162	0.0502**	-0.0147	-0.0317	-0.0126	-0.0390	-0.0204	0.0478*	0.0189	-0.00136
	(0.94)	(3.00)	(-0.44)	(-0.97)	(-0.26)	(-0.81)	(-0.81)	(1.96)	(0.76)	(-0.06)
2008 × OSS	0.247★	0.268★	0.226★	0.244★	0.241*	0.366**	0.288***	0.342★	0.236★	0.254★
	(6.40)	(6.65)	(4.51)	(4.76)	(2.38)	(2.64)	(3.44)	(3.87)	(5.35)	(5.55)
2009 × OSS	0.130★	0.154★	0.0960**	0.117**	0.164*	0.232**	0.202***	0.232★	0.146★	0.152★
	(4.93)	(5.79)	(2.69)	(3.15)	(2.18)	(3.01)	(3.65)	(4.49)	(4.63)	(4.84)
2010 × OSS	0.166★	0.182★	0.134★	0.146★	0.263★	0.292★	0.211★	0.205★	0.184★	0.176★
	(6.60)	(7.20)	(3.91)	(4.10)	(4.02)	(4.19)	(3.99)	(4.19)	(6.15)	(5.90)
2011 × OSS	0.165★	0.166★	0.147★	0.127***	0.213***	0.312★	0.171**	0.167***	0.184★	0.161★
	(6.71)	(6.68)	(4.35)	(3.59)	(3.32)	(4.59)	(3.29)	(3.46)	(6.23)	(5.44)
2012 × OSS	0.146★	0.145★	0.130★	0.113**	0.186**	0.248★	0.166**	0.137**	0.164★	0.143★
	(5.96)	(5.85)	(3.89)	(3.24)	(2.94)	(3.68)	(3.22)	(2.87)	(5.59)	(4.87)
2013 × OSS	0.129★	0.158★	0.119***	0.139★	0.198**	0.246★	0.129*	0.103*	0.147★	0.161★
	(5.28)	(6.41)	(3.56)	(4.00)	(3.16)	(3.68)	(2.50)	(2.17)	(5.04)	(5.53)
2014 × OSS	0.102★	0.128★	0.0964**	0.122***	0.153*	0.215**	0.111*	0.0657	0.121★	0.137★
	(4.16)	(5.19)	(2.89)	(3.51)	(2.44)	(3.22)	(2.15)	(1.38)	(4.14)	(4.69)
2015 × OSS	0.110★	0.133★	0.113***	0.119***	0.164**	0.223***	0.109*	0.0704	0.125★	0.138★
	(4.53)	(5.40)	(3.39)	(3.41)	(2.62)	(3.33)	(2.12)	(1.48)	(4.27)	(4.72)
2016 × OSS	0.0863***	0.115★	0.0749*	0.107**	0.131*	0.208**	0.0844	0.0580	0.0906**	0.115★
	(3.53)	(4.66)	(2.24)	(3.05)	(2.08)	(3.11)	(1.64)	(1.22)	(3.09)	(3.95)
2017 × OSS	0.0688**	0.0951***	0.0594	0.0913**	0.120	0.187**	0.0733	0.0460	0.0717*	0.0991***
	(2.81)	(3.85)	(1.77)	(2.61)	(1.90)	(2.78)	(1.42)	(0.96)	(2.44)	(3.38)
2018 × OSS	0.0521*	-0.0657**	0.0524	-0.0149	0.110	0.00993	0.0408	-0.116*	0.0669*	-0.0255
	(2.11)	(-2.61)	(1.55)	(-0.42)	(1.73)	(0.15)	(0.78)	(-2.39)	(2.25)	(-0.85)
Constant	0.0109	-0.0389**	-0.402★	-0.462★	-0.331★	-0.214★	0.316★	0.240★	-0.358★	-0.354★
	(0.72)	(-2.63)	(-13.94)	(-16.37)	(-7.90)	(-5.12)	(13.60)	(10.73)	(-16.30)	(-16.83)
Student FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	3058431	3048242	776476	776440	442401	442499	1594951	1583522	1588048	1592976
Schwarz Critical Value	3.864	3.864	3.683	3.683	3.606	3.606	3.780	3.778	3.779	3.779

Notes. Estimates for Equation 3 with controls for school composition and student, school, and year FEs. Odd-numbered columns use reading score as the desired output; even-numbered columns use math as their output. The first and second rows display estimates for θ_t and π_t, respectively, from Equation 3 for $t = 2019$. Effects for other years can be obtained by summing the estimate in $t = 2019$ and the corresponding estimate on the interaction term between year and the covariate of interest. t statistics using standard errors clustered at the student level are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ★ $t > \sqrt{\log(N)}$.

Appendix 6. Dynamic model of OSS on test score

	$t-2$		$t-1$		Past		$t+1$		$t+2$		Future		All	
	Reading (1)	Math (2)	Reading (3)	Math (4)	Reading (5)	Math (6)	Reading (7)	Math (8)	Reading (9)	Math (10)	Reading (11)	Math (12)	Reading (13)	Math (14)
OSS_{t-2}	-0.00512*	-0.0157★			-0.0200★	-0.0373★							-0.0294★	-0.0382★
	(-2.56)	(-8.63)			(-9.59)	(-19.58)							(-12.71)	(-18.45)
OSS_{t-1}			-0.0251★	-0.0361★	-0.0293★	-0.0473★							-0.0435★	-0.0545★
			(-15.31)	(-23.85)	(-15.52)	(-27.16)							(-20.06)	(-27.59)
OSS_t					-0.0568★	-0.0774★					-0.0731★	-0.0894★	-0.0730★	-0.0887★
					(-32.70)	(-48.46)					(-47.70)	(-62.76)	(-34.94)	(-46.71)
OSS_{t+1}							-0.0365★	-0.0316★			-0.0426★	-0.0377★	-0.0440★	-0.0413★
							(-27.20)	(-25.17)			(-29.89)	(-28.39)	(-22.32)	(-23.15)
OSS_{t+2}									-0.0151★	-0.00859★	-0.0258★	-0.0205★	-0.0237★	-0.0210★
									(-11.48)	(-6.93)	(-19.17)	(-16.12)	(-12.97)	(-12.68)
% Econ. Disadv.	-0.0346★	0.0323★	-0.0426★	0.0115*	-0.0352★	0.0313★	-0.0404★	0.0180**	-0.0394★	0.00564	-0.0379★	0.00749	-0.0358★	0.0248★
	(-5.27)	(5.11)	(-7.05)	(1.98)	(-5.37)	(4.97)	(-6.99)	(3.19)	(-6.52)	(0.98)	(-6.27)	(1.30)	(-5.08)	(3.80)
% Black	0.114**	-0.0988**	0.0804**	-0.118★	0.118**	-0.0934**	0.0388	-0.0787**	0.0416	-0.0681**	0.0444	-0.0654**	0.136***	-0.0582
	(2.99)	(-2.73)	(2.63)	(-4.03)	(3.10)	(-2.58)	(1.52)	(-3.13)	(1.58)	(-2.68)	(1.69)	(-2.58)	(3.36)	(-1.57)
% Hispanic	0.312★	0.0469	0.250★	0.101**	0.321★	0.0607	0.185★	0.130★	0.192★	0.167★	0.195★	0.169★	0.340★	0.130**
	(7.75)	(1.22)	(7.68)	(3.22)	(7.97)	(1.58)	(6.77)	(4.80)	(6.82)	(6.10)	(6.93)	(6.21)	(7.91)	(3.28)
% White	0.154★	-0.127***	0.129★	-0.0695*	0.158★	-0.121***	0.0968★	0.00884	0.0974★	0.0312	0.102★	0.0354	0.164★	-0.0649
	(4.29)	(-3.69)	(4.44)	(-2.49)	(4.39)	(-3.53)	(3.97)	(0.37)	(3.88)	(1.29)	(4.05)	(1.46)	(4.28)	(-1.84)
Constant	-0.119***	0.0823**	-0.0847**	0.0612*	-0.114***	0.0888**	-0.0504*	0.0000410	-0.0542*	-0.0125	-0.0483*	-0.00577	-0.112**	0.0574
	(-3.69)	(2.68)	(-3.27)	(2.46)	(-3.55)	(2.90)	(-2.32)	(0.00)	(-2.42)	(-0.58)	(-2.16)	(-0.27)	(-3.26)	(1.82)
Student FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	1982418	1966123	2519501	2506811	1982415	1966120	3021626	3012225	2878360	2894459	2878358	2894457	1822704	1832753
Schwarz Critical Value	3.807	3.807	3.839	3.839	3.807	3.807	3.863	3.862	3.857	3.857	3.857	3.857	3.797	3.798

Notes. Estimates of the regression specified as $y_{it} = \sum_{k \in \mathcal{T}_m} (\alpha_k \text{OSS}_{i,t+k}) + \varphi W_{it} + \xi_i + \lambda_s + \delta_t + \varepsilon_{it}$ for $m = 1, \ldots, 7$, where $\mathcal{T}_1 = \{-2\}$, $\mathcal{T}_2 = \{-1\}$, $\mathcal{T}_3 = \{-2, -1, 0\}$, $\mathcal{T}_4 = \{1\}$, $\mathcal{T}_5 = \{2\}$, $\mathcal{T}_6 = \{0, 1, 2\}$, and $\mathcal{T}_7 = \{-2, -1, 0, 1, 2\}$. Controls for school compositional variables and student, school, and year FEs included. Odd-numbered columns use reading score as the output; even-numbered columns use math score as the output. t statistics using standard errors clustered by student are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ★ $t > \sqrt{\log(N)}$.