Whose IDEA Is This? A Case Study Examination of the Effectiveness of Inclusive Education

Katharine Parham Malhotra[†]
Teachers College, Columbia University

March 2023

Abstract

The inclusion of students with disabilities in general education settings has steadily increased since the 1990s. Yet little is known about whether inclusive education is effective for these students or their nondisabled peers. I examine the impacts of inclusive education on both student groups through the lens of a school district that implemented a policy of inclusion as the default student placement in the early 2000s. I leverage the staggered, school-level implementation in an event study model to estimate the policy's impacts on academic and behavioral outcomes. I find there were no detrimental impacts on the academic performance of students with or without disabilities as a result of the policy. Elementary and middle school students' standardized test scores, as well as attendance rates across all school levels, were unaffected. High school graduation and 9th grade promotion rates were unaffected during implementation but increased by two and six percentage points, respectively, over the subsequent five years. This study offers evidence that inclusive education does not come at the expense of academic progress for students with or without disabilities in the short term, and may improve some academic outcomes in the longer term.

[†] Ph.D. Candidate in Education Policy, Teachers College, Columbia University, Department of Education Policy and Social Analysis. Email: katharine.parham@tc.columbia.edu.

1 Introduction

Since the introduction of compulsory schooling in the 19th century, stakeholders have debated how to best educate children with special needs. A public awakening to discrimination issues in the 1960s, along with the 1975 passage of federal special education law, helped transition the U.S. from two, segregated forms of schooling to one in which students with disabilities were not considered inherently different (Winzer, 2012). More than a half century after the passage of the Individuals with Disabilities Education Act (IDEA), there is little conclusive evidence on the most effective ways to educate this vulnerable population. This study examines one, anonymous U.S. school district whose transition to full inclusivity over an eight-year period in the early 2000s allows for drawing broader conclusions about one of the most pervasive challenges in education—managing the diverse, individualized needs of all students within a classroom. I leverage the staggered policy adoption within the district and an event study approach to estimate the policy's impacts on the academic and behavioral outcomes of students with and without disabilities.

IDEA requires that states develop procedures to ensure students with disabilities are educated, to the greatest extent appropriate, alongside peers without disabilities in students' least restrictive environment (LRE). Though the law never uses the term "inclusion," advocates and practitioners have interpreted the motivation of the LRE mandate as including as many students as possible in their local community school, inside a regular, grade-level-appropriate classroom for as much of the day as possible (Dorn, Fuchs, & Fuchs, 1996; Giordano, 2007). This practice of prioritizing the placement of students with and without disabilities in the same learning environment has become increasingly prevalent in recent years (see Figure 1), and today more than 60% of all students with disabilities nationwide spend 80% or more of their day in general education environments—up from just 30% in the early 1990s (NCES, 2019a).

While this practice has become increasingly popular over the past 30 years, its use is not supported by a robust or coherent evidence base, and even its proponents disagree on the merits. Some supporters of inclusive education argue on largely ideological grounds, seeking course correction from an insidious, segregated history of educating children with special needs (Crockett, 2020). Others argue inclusive settings benefit students with and without disabilities, emphasizing the cognitive and non-cognitive benefits for all students from time spent learning in a "diverse" environment (Sanger, 2020; Peltier, 1997; Salend & Duhaney, 1999). Empirical research on the effectiveness of inclusive settings for students with and without disabilities is similarly conflicted.

While some research suggests students with disabilities educated in inclusive settings are more likely to make academic progress and graduate on time (Dessemontet & Morin, 2012; Schifter, 2015), other work has found the impact of educating students with disabilities in the general education classroom to have adverse effects both for them (Daniel & King, 1997) as well as their peers without special needs (Fletcher, 2010; Gottfried, 2014).

This paper proceeds as follows. Section 2 anchors this study within the body of extant knowledge on inclusive education. It also provides details on the policy context and how inclusion was implemented. Section 3 describes the data used and methodological plan for the event study analysis. In Section 4, I present results while Section 5 describes a series of robustness checks. Section 6 concludes with a discussion of the findings and recommendations for future research.

2 Background

2.1 What do we know about inclusion?

IDEA was most recently reauthorized in 2004 and did not prescribe one path for all children with disabilities, but rather created a process by which a team of individuals who know a child can best determine what is appropriate for the child's education. The four basic provisions of IDEA ensure that, regardless of a child's unique needs: 1) they are entitled to an appropriate education at the public expense; 2) a continuum of placements must be available to every student with a disability; 3) every student will be educated in their least restrictive environment (LRE); and 4) every student with special needs will have an individualized education program (IEP) providing for those needs (IDEA, 2004). The third provision, describing the placement of students with disabilities in the appropriate educational environment, speaks directly to the role of the least restrictive environment and is the most relevant to the present study.

Federal regulations mandate that states monitor the implementation of this LRE provision and annually report the proportion of time school-aged students are educated in the general education classroom across four main categories:[1] 1) more than 80% of the school day; 2) 40 to 79% of the school day; 3) less than 40% of the school day; or 4) all of the school day in a separate

[1] The first three categories were originally: 1) more than 60%, 2) 21 to 60%, and 3) less than 21%, up until 2008. For the purposes of this study, and for continuity of analysis before and after this data-collection shift, these three categories have been recoded in this study's data to reflect the newer percentage bands for the years prior to 2008. This does not affect the number of students reported as included, but it is notable that the thresholds for inclusivity were lower prior to this date.

setting.[2] Figure 2 illustrates the full continuum of service provision as it relates to these categories and illustrates the policy's intent—that the majority of students should be in general education settings for the majority of the school day (the least restrictive setting) and only a small number of students should be spending most of their time in isolation. The first of the four reporting categories, reflecting students spending more than 80% of the school day in the general education classroom, is synonymous with the idea of inclusion, though no federal laws or regulations offer an explicit definition of the term, and preferred terminology to describe the same concept—children with special needs educated primarily within the general education classroom—has evolved over time.[3]

Inclusion, while ill-defined, is also difficult to rigorously examine. The lack of consistent definition means inclusion may be implemented completely differently from one context to the next. There are also empirical challenges, consistent with those in the broader literature on special education effectiveness. Students with disabilities do not have an obvious comparison group among peers without disabilities, and examining students with disabilities among themselves is limited by issues of selection into special education and the differences across individual students. Some studies have attempted analyses of the causal impacts of special education by examining within-student variation; that is, examining the academic performance of students who enter and exit special education over their educational careers (Hanushek et al., 2002), though this approach remains limited by selection issues. On average, both empirical and observational evidence suggests that students who are identified for and receive special education services have improved test scores (Hanushek et al., 2002; Rea et al., 2002; Schwartz et al., 2019) and long-term educational attainment (Ballis & Heath, 2019), though some evidence using matching methods has found special education generally to have a negative or insignificant impact on identified students' learning and behavior (Morgan et al., 2010).

Research on inclusion, specifically, is also often limited to observational methods; however, these studies still offer useful conclusions for specific contexts and directions for future

[2] The data for this fourth category is maintained at a more granular level, as well (e.g., tracking the number of students in hospital settings, nonpublic school placements, residential treatment facilities, etc.).

[3] Notably, though still not defined, inclusion is also emphasized in other pieces of federal legislation. The Workforce Innovations and Opportunities Act (WIOA) and the Workforce Investment Act (WIA) both include language in their most recent reauthorizations promoting integrated outcomes for youth with significant disabilities and increasing accountability for schools in ensuring that inclusive workforce opportunities are provided (Workforce Innovation and Opportunities Act, 2014; Workforce Investment Act, 2014).

research. On average, extant observational studies suggest that when students with disabilities are included in general education, their outcomes improve even when controlling for peer, school, and district characteristics (Schifter & Hehir, 2018; McLeskey, Rosenberg, & Westling, 2018). This is true for both academic and non-cognitive outcomes, as evidence of improved test scores is often observed alongside improved work habits, self-confidence, social competence, and attentive behavior (McLeskey, Rosenberg, & Westling, 2018). There is also evidence that inclusive education results in null or insignificant effects for students with disabilities (Affleck et al., 1988; Jenkins et al., 1991).

Studies on inclusion have also examined how students *without* disabilities in general education fare in inclusive settings. Research examining the peer effects associated with inclusive practices suggests largely negative effects on students without disabilities, with the important caveat that many studies focus exclusively on the impacts of learning alongside students with significant behavioral problems—an attribute not representative of all students with disabilities. Exposure to classmates with disruptive behaviors has been shown to have negative academic effects on other students in terms of both math and reading test scores (Fletcher, 2010). Peer behavior is similarly affected, with increases in the number of classmates with disabilities associated with lower levels of self-control and interpersonal skills among students without disabilities (Gottfried, 2014), as well as a potential reduction in lifetime earnings (Carrell, Hoekstra, & Kuka, 2016).

Evidence from inclusion studies which do not focus on the behavior of students with disabilities has found a mix of negative (Robinson, 2012), positive (Sharpe, York, & Knight, 1994), and null (McDonnell et al., 2003; Brewton, 2005; Brady, 2010; Trabucco, 2011) impacts of inclusive education on the academic performance of students without disabilities. The variation in findings again suggests that the specifics of how inclusion is implemented in a given context matters significantly for observed outcomes. However, the confluence of evidence when the behavior of students with disabilities is not the primary independent variable suggests the academic performance of students without disabilities is often unaffected by the increased presence of peers with special needs in the same classroom environment.

Much of the conflicting evidence on inclusion can be attributed to the lack of clear definition and the implementation differences across contexts. Issues of students' access to grade-level curriculum within the classroom, levels of individualized supports available, and the manner

in which teaching practices necessarily change when students with disabilities are included further limit understanding of the specific mechanisms underlying the effectiveness of inclusive education. Evidence on co-teaching—a common practice for implementing inclusion—suggests the staffing strategy has positive academic impacts on students with and without disabilities in inclusive settings (Tremblay, 2012; Jones & Winters, 2020), but more work is necessary to understand why this may be the case. A final caveat is that not all students with disabilities make progress in inclusive settings, even if performance improves on average, and students with different disability classifications cannot be treated interchangeably (Gilmour & Henry, 2018; Schulte & Stevens, 2015). Students with low-incidence, or severe, disabilities are disproportionately placed in more restrictive settings (Smith, 2007; Kurth et al., 2014), limiting knowledge of how students with the most significant needs may fare in inclusive environments.

2.2 Policy Context

Setting

The school district represented in this study is anonymous. However, descriptive details contextualize the setting in which this policy transition took place, for exposition and generalizability considerations. Table 1 describes the student population enrolled in the case study district. Notably, the majority of students enrolled are white, non-Hispanic. Approximately 16% of students qualify for special education—two percentage points higher than the national average for public schools (NCES, 2019a). Just under half of all students qualify for free or reduced-priced lunch, and less than three percent of students are English-language learners. The district is located in a rural region, less than five miles from an urbanized area, with a population of less than 20,000 students. Roughly 10% of families in the district fall below the poverty line, and 80% of households have access to the internet (NCES, 2019b). The rural setting of this case study district is notable, as rural districts face particular special education challenges including issues of teacher retention and recruitment and transportation (Helge, 1981), and more than half of all school districts in the U.S. are located in rural environments (AASA, 2017).

Status Quo

Prior to policy implementation, students with disabilities in the case study district were largely segregated from their non-disabled peers. While students without disabilities were served

almost exclusively in their neighborhood schools, students with disabilities—16.8% of the district's total student population—were served in a specialized environment. Two schools in the district served as "centers" in the pre-policy period, with targeted programs for specific student populations, including those with significant cognitive impairments, emotional disturbance/behavior disorders, visual or hearing impairments, and autism. 10.7% of students with disabilities received special education services in one of these centers and nearly 100% of those students rode specialized school buses to and from these locations. Seven schools in the district also offered self-contained programs for students with severe cognitive impairments and emotional disabilities, while all district schools offered in-school resource classrooms for pull-out services. Prior to the policy, the case study district had one of the highest placement rates of students with IEPs in non-inclusive settings among districts within the state, with 49% of all students with disabilities educated for the majority of their school day in separate classes, fully segregated settings or regional centers.

In addition to high rates of separation, academic performance of students with disabilities in the district was among the lowest statewide in the years preceding the inclusion policy. A separate, "parallel" curriculum was used in segregated classrooms, meaning students with disabilities received content distinct from their general education peers in the same grade levels. General education and special education teachers also received separate professional development (PD) programs, with special educator development focused on process and legal issues while general educator development addressed content and student achievement indicators. Both student placement patterns and the separation of professional development in the pre-policy period contributed to a lack of collaborative opportunities for instructional staff and the continuation of segregated educational plans for students with disabilities.

In the years prior to the policy, the county in which the district is located also experienced a steady influx of students with disabilities from surrounding areas, both in and out of state, given its geographic location near the state border and a reputation for offering a large number of specialized services. This influx resulted in an associated increase in the costs of special education service provision, on top of already high district costs proportionate to the size of the student population being served. Transportation costs were a particular pain point, with a large proportion of students with disabilities requiring specialized busing across the district to non-neighborhood schools. While the county in which the district is located is small in terms of population (and the

associated tax base), it is large by geographic area, which increases costs particularly for transporting students to schools to which they do not live in close proximity. Roughly 20% of the district's total annual transportation budget was allocated to special needs transportation in the years prior to the policy, with annual costs per special education student at nearly $4,000, compared to roughly $500 per general education student. These existing financial concerns, alongside concerns about low student performance, drove district leaders to reexamine their approach to special education and consider ways to increase districtwide inclusivity as a potential solution.

2.3 Policy Implementation

Process

For support with implementation, the case study district sought expertise from a nonprofit organization with experience facilitating whole-system transformation centered around inclusive practices.[4] The organization had experience working with other districts in and out of state and provided staffing, professional development, and technical assistance for the district staff throughout implementation.[5]

The transition to inclusion at each district school followed a four-year arc, at the end of which schools would be, and remain, fully inclusive. The district created cohorts of four to eight schools and arranged them into a predetermined order for each year of the eight-year transition period. Though there was some variation, in general, district implementation began with elementary schools, followed by middle schools, and then high schools. Within school levels, the order of schools implementing the policy was close to random; that is, no specific criteria (e.g., test scores, stated willingness, size of special education population) were used to determine the order of implementation among elementary, middle, or high schools. Figure 3 shows the district's eight-year implementation plan. Individual, school-level transitions over the four-year arc followed a consistent process, designed as a gradual-release model in which the capacity of each school slowly increased alongside a decrease in support from the external partner.

[4] Given the extent of the relationship between the nonprofit and the district in facilitating the implementation, it would be difficult, if not impossible, to retroactively disentangle the implementation effects of the policy from the effects of the district having worked with this particular nonprofit organization. This study does not attempt to tease apart these interrelated, but distinct, strands of influence on the policy's effects. However, future work examining similar policies—particularly evaluations taking place in real-time—could better examine each stream of influence.

[5] None of the work by this entity has been previously rigorously examined by external evaluators.

In the first year of the policy transition, the focus for each school was exclusively on planning. The policy itself, while designed to facilitate systems change, was also largely student-centered, and emphasized transitioning students with disabilities from non-neighborhood schools into general education within their neighborhood community schools, as well as bringing students from more segregated settings within their community schools into general education. In the first year, schools worked to identify specific, student-level needs and planned for individual students to transition.[6] During the first year of the transition, schools also began participating in professional development administered by the nonprofit partner—roughly 8 hours of inclusion-specific development over the course of the school year. During these sessions, schools participated in needs assessments to self-identify areas of growth within their knowledge and practice and to develop shared visions of what inclusion would look like within their unique campus.

In the second year, individualized student transitions began in earnest. Teams from both the receiving and sending schools met one-by-one with the families of each special education student, and the students themselves when age-appropriate, in a series of meetings to discuss the transition process and plan individualized support structures to ensure student success in the general education classroom in the receiving school. A minimum of two meetings per student took place prior to a student transition.

The third year of implementation focused on developing and solidifying whole-school structures to support inclusive education, such as schedule revisions to allow for collaborative planning, and included additional professional development on best practices for collaborative teaching. Whole-school revisions began by identifying student needs, then considering the staff required to facilitate those needs and how to allocate existing staff strategically. The structure of staffing roles was dependent on needs. One special education teacher might be in a traditional, co-teaching relationship with one general education teacher for reading every day, and in a consultative relationship around student behavior with another general educator. Once staff allocations and role structures were assigned, the master schedule was developed such that planning periods were shared for all individuals whose roles required regular collaboration.

[6] The nonprofit partner encouraged the district to aim to transition 100% of students with disabilities into general education classrooms within their neighborhood schools, with the caveat that all families had the right to decline a change in least restrictive environment placement for their child.

The fourth and final year of each school's transition emphasized improvements to the quality of instruction in classrooms and the meaningful participation of all students. The nonprofit's role during this final year was primarily consultative on any outstanding issues. Professional development during this year was not predetermined by the nonprofit partner, but was designed to be responsive to the outstanding needs and challenges faced by schools in their final year of implementation. Following the fourth year, direct, school-level support from the nonprofit partner concluded.

Implementation Success

Before examining the policy's impact, I first establish it was successfully implemented. Figure 4 provides evidence of this success, showing the case study district's rate of inclusivity over time, as measured by the percentage of students with disabilities placed in the general education setting as their primary learning environment. This figure demonstrates that the district steadily increased the number of students with disabilities spending 80% or more of their day in general education settings over the course of the eight-year policy transition period, increasing from a district average of 60% of students with disabilities just prior to the policy transition to more than 90% in subsequent years. This high rate of inclusivity sustained in the decade following the end of the implementation period.

Figure 5 compares the inclusivity trends of the case study district to all other districts in the state over the same twenty-year period, showing the same increase during the implementation period and sustained, high rates of inclusion in subsequent years. This figure also demonstrates that while many districts in the state followed the national trend of increased inclusion over this period, inclusion rates in the case study district moved beyond those observed elsewhere.

3 Data and Methods

3.1 Data

I use nonpublic, school-level data from the case study district, along with publicly available data on academic outcomes from the district's associated state department of education and the National Center for Education Statistics' (NCES) Elementary and Secondary Information System (ELSI), and data on behavioral outcomes from the U.S. Civil Rights Data Collection (CRDC). I construct an original panel dataset that allows observation of the key independent variable—

placement into general education compared to other educational environments along the continuum of service provision—along with both academic and behavioral measures of effectiveness.[7] Academic outcomes include high school graduation and dropout rates, rates of grade promotion and retention, and performance on state standardized reading and math assessments. Behavioral outcomes include attendance rates and the discipline of students with and without disabilities, including instances of suspensions, expulsions, and the use of corporal punishment.[8]

The main academic measures rely on standardized test score data reported by the case study state's department of education. Like many states following the introduction of No Child Left Behind in 2002, student performance is annually reported for all students and student subgroups on reading and math assessments in grades 3 through 8 as the percentage of students in a given grade in a given year performing at a proficient level. States do not report school or district test score means, and under NCLB, states were allowed to determine their own thresholds for proficiency—presenting challenges for both long-term measurement and analysis.[9]

Following the work of Reardon, Shear, Castellano, and Ho (2016), I use homoskedastic ordered probit (HOMOP) to transform the reported frequencies of students scoring proficient or above on annual standardized reading and math assessments in grades 3 through 8 into estimated means and standard deviations.[10] While heteroskedastic ordered probit models enable the transformation of student frequencies across multiple performance groups (e.g., students performing at basic, proficient, and advanced levels; or, performance levels ranging from one to

[7] There are a number of outcomes for which data are not available that would provide a deeper understanding of this policy's effectiveness. Data on socioemotional skill development, students' mental health, and more robust indicators of student behavior, among others, would each be highly relevant to future examinations of the impacts of inclusive education.

[8] Disciplinary data are only available for some years, given the staggered intervals over which CRDC data are collected and a change in the specifics of data collected during the policy implementation period analyzed in this study. Therefore, disciplinary data are used only for descriptive analysis in the post-implementation period rather than the main causal analysis.

[9] While each state determines the cut points along the full spectrum of potential raw scores for what qualifies as performing at basic, proficient, and advanced levels on their state exam—often for arbitrary and political reasons (Ho, 2008), and making cross-state comparisons impossible—state-set cut points may also change from year to year depending on bureaucratic assessments of the difficulty of each year's exam. Changing thresholds of proficiency on an annual basis makes it difficult to reliably compare student performance over time using proficiency rates alone, even within a single state.

[10] The HOMOP model is a variation on the heteroskedastic ordered probit (HETOP) model used by Stanford researchers to construct the Stanford Education Data Archive (SEDA). SEDA uses only state-reported proficiency rates to compile a dataset of student achievement that is reliably comparable across states and over time, despite the limitations of the proficiency metric.

five) and therefore across multiple cut points, the HOMOP model is better suited to instances with just one cut point. The data available for this study is just that: a reporting of the number of students who performed above or below the set proficiency threshold each year.

I use a HOMOP model to estimate a unique mean for each student group (all students, students with disabilities, and students without disabilities) on each reading or math assessment within each school in each year for which data are available. Each subgroup's subject-by-grade-by-school-by-year estimate is transformed from a frequency into an inference of that subgroup's *propensity* for proficiency. Under the assumption that test score distributions are normal, probit allows for a transformation of percents-proficient into standard deviation units, which are implied differences in averages. This rescaling corrects for potential distortions that occur when proficiency thresholds are set near the extremes of normal distributions.[11] One constraint of having only two proficiency categories is an assumption that all group variances are assumed as equal, and therefore the standard deviations for each subject-by-grade-by-school-by-year estimate are set to a single, fixed constant of 1 (Fahle et al., 2017). The resulting means for each group, however, are unique. A second constraint is that, as a result of the nature of HOMOP transformations, resulting means for students with and without disabilities are only comparable *within* and not *across* subgroups, limiting the ability to speak to achievement gaps directly.

This transformation approach is also limited in the cases of insufficient data or small sample sizes. Reardon and colleagues demonstrate that accurate estimations of means and standard deviations of test score distributions are possible, particularly when sample sizes are larger than 50 (2016).[12] Imprecise estimates are more likely when sample sizes fall below 50, or if underlying distributions are not normal. This primarily affects a key subgroup of interest in this study—students with disabilities—for whom within-school frequencies are naturally small. 85-98% of school-level proficiency counts (variation across subjects and grades) for students with disabilities in the data used in this study are below this threshold and therefore have estimated test score means

[11] As an example, if a proficiency threshold is set at the 50th percentile (the top of a normally distributed bell curve), many students will naturally perform around that marker and purely mathematically more students will have opportunities to cross the threshold for proficiency than if the threshold were simply set closer to one of the extremes on the distribution, where there are fewer students. More students crossing a 50th percentile threshold for proficiency will look like a larger percentage of an overall population making meaningful gains than if a smaller number of students were to cross a threshold set much higher along a normal bell curve.

[12] The authors demonstrate this by comparing estimated means and standard deviations in instances where both proficiency rates and raw test score averages are available and find they are able to retrieve reliable estimates of group means with this method.

that are potentially slightly negatively biased. However, Reardon et al. note that even when sample sizes are small, average bias is not sizable with respect to the true standard deviations in the underlying data (2016).

Figures 6 and 7 illustrate the impact of these transformations on the underlying proficiency data in this study for reading and math assessments. Both figures show average student performance by district over time, with the pre-transformation trends in column 1 and the transformed performance data in column 2. Trends are broken down for all students (row 1), students with disabilities (row 2), and students without disabilities (row 3) across all assessments in grades 3 through 8. Figure 6 shows this transformation comparison for math assessments, while Figure 7 reflects the trends for reading. For both subjects, and across all student groups, in the pre-transformed data there is an apparent, steady increase for all districts over time up until 2012 (when PARCC began influencing curriculum decisions in classrooms), followed by a subsequent decline. This same trend is not observed in the transformed data. Student performance appears consistent across districts over time, particularly for students without disabilities. The slight year-to-year fluctuations in trends for students with disabilities are likely an artifact of the data limitations resulting from small sample sizes, though student performance for students with disabilities in all districts still hovers around zero. These figures make clear that viewing student proficiency as a rate alone distorts actual trends in student performance. While the percentage of students performing proficient or higher on state exams does steadily increase in the years leading up to the PARCC rollout, actual student performance in terms of estimated means is more consistent. These retrieved test score means are used as the key indicator of academic performance in the event study analysis.

3.2 Methods

To estimate the impact of the district's policy of inclusion on students' academic and behavioral outcomes, I use a variation of the standard two-way fixed effects (TWFE) difference-in-differences (DiD) strategy that draws on variation in the year in which a school began implementing the inclusion policy, allowing for examination of potentially dynamic treatment effects. The following model, accounting for this staggered adoption of treatment, reflects the main estimation of the policy's effectiveness:

$$y_{st} = \propto_s + \ \delta_{tg} + \sum_{\substack{k=-3 \\ k\neq -1}}^{k=8} 1(t = t_s^* + k)\beta_k + \ \varepsilon_{st}$$

In this specification, y_{st} reflects either an academic or behavioral outcome of interest for a given school, *s*, in a given year, *t*. The parameters $\propto_s$ and δ_{tg} indicate the inclusion of both school and year-by-school-level fixed effects, respectively, controlling for school-invariant and time-by-school-level-invariant differences across schools. The latter restricts within-year comparisons to schools at the same level (e.g., elementary, middle, high). The effect of the policy implementation beginning in year t_s^* is reflected in the coefficient β_k, relative to outcomes k years later. The model traces out the comparison between treated and untreated schools from three years prior to policy implementation for a given school to eight years after implementation began, omitting the year prior to the start of implementation as the excluded group. The variation that identifies each β_k therefore comes from the interaction between within-school changes and time, as two comparisons of the outcome variable: 1) comparing to the years before the policy change began for a given school, and 2) comparing treated and untreated schools within the same level and academic year.

Given that policy implementation for each school began in a given year, continued over four years, and then remained "on" in perpetuity, this estimation strategy allows for observation of the policy's potentially heterogeneous effects throughout the formal treatment period as well as after treatment has concluded but when schools remain effectively treated. In short, the event study allows for an understanding of the policy treatment from beginning to end, explicitly modeling the dynamic treatment effects across time. In all models, standard errors are clustered at the school level to address any potential bias resulting from serial correlation across outcome variables given that data span multiple years and variation occurs only at the group level (Angrist & Pischke, 2015). Conventional standard errors would be biased downward.

Results from the event study are presented graphically. I also estimate a piece-wise spline function where I decompose a more traditional DiD estimate into an "implementation period" (years 0-4) and a "post-implementation" period (years 5+). This model captures the most important differences in policy response and allows for variation in impact over time, but has greater statistical power than the event study. Four design choices and assumptions underlie the causality

of findings from this model: comparison group selections, parallel trends, exogenous assignment to treatment, and homogeneous treatment effects.[13]

Comparison Group Selection

I use all other untreated schools within the state as the comparison group. The large number of untreated schools offers power, and the high level of certainty about their having never received treatment lends substantive confidence to the choice. The limit of this approach is that the sample of schools statewide differs from the schools in the case study district in terms of both demographics and geography, as the larger sample necessarily includes a wider range of school sizes, compositions, and locations (see Column 2 of Table 2).

I also consider limiting the comparison group to only schools located in other rural districts within the state, as the case study district is in a rural setting. Given that rural school districts face particular challenges with respect to special education, this approach seemingly has substantive merit. However, there remain both demographic and geographic differences between the populations of schools in these untreated, rural districts and those in the case study district, despite their shared "rural" indicator (see Column 3 of Table 2). Additionally, the diminished sample size results in a substantive loss of power.

A synthetic comparison group is a third option—a subset of schools drawn from the full pool of untreated schools within the state based on a set of observable characteristics used to match to the set of treated schools.[14] Descriptive statistics for this third group are shown in Column 4 of Table 2 and are highly similar to those of the full population of untreated schools within the state. As such, the main results are based on a comparison to the broader comparison group, though to

[13] Literature on the use of both TWFE DiD models as well as event study designs has expanded considerably in recent years (Goodman-Bacon, 2020; Athey & Imbens, 2021; Callaway & Sant'Anna, 2020; De Chaisemartin & D'Haultfœuille, 2018; Roth, 2019; Sun & Abraham, 2021; Sant'Anna & Zhao, 2020), casting new light on previously unquestioned assumptions about the DiD approach.

[14] The synthetic comparison group is generated using a propensity-score matching technique, radius matching with replacement, allowing for multiple matches for each treated unit. Estimates from this matching method are more precise than one-to-one nearest neighbor or nearest neighbor without replacement matching alternatives given the resulting, larger sample size (Somers et al., 2013)—particularly important in this instance where the number of treated units is small and maximizing sample size is a key concern. Covariates for school-level propensity score prediction include: school location (e.g., urban or rural), size of school (total enrollment), proportion of students in special education, proportion of students receiving free or reduced-price lunch, and the proportion of students in each major racial/ethnic subgroup. A radius of 0.01 was used to create the final, synthetic sample.

assess the sensitivity of findings to this choice, results compared to both comparison group alternatives are presented in an Appendix.[15]

Parallel Trends

I compare demographics of schools in the case study district to those in all other untreated districts in the state to test for the presence of parallel trends. Data on key outcomes in the case study district are not available until the first year of policy implementation and render outcome-based pre-trends unobservable. Inability to observe pre-trends is a common limitation across difference-in-differences studies (Roth, 2019); however, historical demographic data are available to assess pre-trends and offer some evidence that there were no substantive changes to the case study district or comparison districts over the arc of the policy implementation period.

Demographic data from 1986-2019 offer 16 years of pre-trend information. As a proxy for outcome data, these data show that based on the composition of case study district schools compared to untreated schools in the state by race, gender, disability status, and the percent of students eligible for free or reduced-price lunch, parallel trends do exist in the period preceding treatment (see Figure 8). This confirms that the composition of the case study district did not substantively change before and after the policy implementation and that the path of untreated schools in the state serves as a meaningful comparison for the case study district.

Exogenous Assignment to Treatment

The nature of the policy implementation was such that all schools were eventually treated, resulting in no never-treated schools. The variation used for identification in this study is the timing of treatment. Causal interpretation of DiD results relies on an assumption of exogenous assignment to treatment—that treated units' assignment to treatment is either random or as-good-as random. Implementation across district schools began with elementary schools, followed by middle and

[15] One other option would be to contain the analysis to just the case study district, but to use the last-treated units as the counterfactual population of schools. In essence, since the path of these schools over the arc of the policy period is unaffected up until the moment their own treatment begins, for the period prior, their path offers a reasonable counterfactual for the schools treated earlier in the panel. There are three downsides to this within-district approach. First, the number of schools within the case study district is small, given its rurality, and reducing the number of treated units would further reduce analytical power. Second, the length of time between the first treated cohort and the last is just four years, offering a relatively limited period over which to observe a hypothetical counterfactual path. Third, given the structure of the policy implementation, the schools who were last to implement the treatment were primarily high schools, and are therefore unlikely to be an appropriate comparison population for the larger group of elementary and middle schools treated in the earlier years.

high schools. While this was not random, the selection of schools *within* levels was as-good-as random; that is, the order was not determined by schools' level of pre-policy effectiveness, openness to inclusion, or some other qualifying criterion.

A threat to validity is if the anticipation of treatment led schools to begin policy implementation at a time other than their assigned start of treatment. Despite schools' knowing the order of implementation in advance, there is no evidence this happened in practice. A key component of the policy required transitioning students with disabilities one by one from the schools in which they were located into their local community schools. This required a series of meetings among the staff at the receiving school, the staff at the sending school, and the families of each student. It is unlikely these stakeholder groups added additional meetings outside the prescribed order, or would have wanted to move individual students to inclusive schools and classrooms without the broader changes to school-wide structures already in place.[16]

"Snowball effects" similarly threaten validity if they disrupt the purity of treatment timing. As an example, 8th graders in the first year of implementation in 2004 would "bring" the policy with them as they entered into a high school in its first year of the policy in 2005. However, all high schools within the district enrolled students from multiple feeder middle schools, and two of the district's six middle schools did not begin their own policy implementation until 2006—the same year as the final cohort, in which all high schools also began treatment. Therefore, in each high school's first year of treatment there were still special education students moving through their first year in inclusive classrooms. Additionally, each high school was implementing their first year of whole-school structural changes in support of the new policy. These changes were necessarily different for high schools than middle schools, so even students familiar with the concept of inclusion would have been adapting to new policies in a new environment. Snowball effects are therefore similarly unlikely to have disrupted the timing of treatment.

Heterogeneous Treatment Effects

A final threat to validity results from drawing on a longer panel of data, wherein early-treated schools become incorporated within the comparison group for later-treated schools, muddling the identification of average treatment effects (Goodman-Bacon, 2020). This is an issue

[16] Conversations with the nonprofit that helped facilitate the policy implementation confirm there were no significant "spillover" effects of the policy to schools prior to their scheduled start of treatment.

if there are differences in the impact of the treatment over time.[17] To establish whether heterogeneous treatment effects are present, I first compute the weights associated with each individual difference-in-differences estimator of average treatment effects (ATT) underlying a standard TWFE regression where differential treatment timing exists.[18] Following de Chaisemartin and d'Haultfœuille (2018) and Goodman-Bacon (2020), identical weights or a lack of negative weights would indicate homogeneous treatment effects over time and no additional steps would be needed. I find that there are negative weights associated with more than one individual TWFE regression and that the magnitude of the weights varies over time—evidence of heterogeneous treatment effects.

Because schools began treatment in "cohorts" of 4-to-8 each year of the implementation period, I next compute the weights of each cohort-specific average treatment effect on the treated (CATT) underlying the TWFE regressions in the event study specification.[19] These weights better reflect the impact of heterogeneous treatment effects over time, in the appropriate context of dynamic treatment effects. Figure 9 displays these weights for each of the five treated cohorts over their respective, four-year policy implementation periods and shows the CATTs of the first and second cohorts are weighted the least and the CATT of the third treated cohort weighted the most over this period. This offers further evidence that there were different policy impacts for each treated group.

I take three steps to address this. First, the nature of the policy implementation was such that the order of treated schools was almost perfectly correlated with the *type* of school (i.e., cohorts comprised groups of elementary, middle, or high schools). It is probable that the initial TWFE weights analysis conflates this strong correlation between treatment timing and school level in the data. In this case, heterogeneous treatment effects are logical given the reasonable expectation that different school levels might respond differently to the inclusion policy. This

[17] Here I reference the issue with treatment effects that differ over *absolute* time. That is, impacts among units who began treatment in one of the first years of the implementation period are different from impacts among those whose treatment began in a later year. This is distinct from the idea of heterogeneity over time since treatment began, or heterogeneity over *relative* time, which I explicitly model through an event study approach to explore potentially dynamic treatment effects.

[18] These estimations were conducted using the `twowayfeweights` command in Stata, which specifies the number and sum of positive and negative weights as well as the measure of robustness to treatment effect heterogeneity in two-way fixed effects regressions (de Chaisemartin & d'Haultfœuille, 2018).

[19] Following Sun and Abraham (2021), these weights are calculated through an auxiliary regression depending only on the distribution of cohorts and indicators of relative time, using the `eventstudyweights` command in Stata (Sun, 2020).

expectation is further corroborated by the CATT analysis illustrated in Figure 9. Given this, I present school-level-specific results along with the main analyses aggregating the policy's impacts for all treated schools. Additionally, in all models I control for year-by-school-type fixed effects. As a final step, results are also estimated using an "interaction-weighted" (IW) estimator that reflects a weighted average of each cohort-average treatment-on-the-treated estimate.[20]

4 Results

Results from the main event study are presented in Tables 3 to 6, with findings reflected for all students in all grades (3-12), elementary school students (grades 3-5), middle school students (grades 6-8), and high school students (grades 9-12), respectively. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded).[21] The pre-policy period is omitted as a reference group. This piece-wise spline specification allows for different slopes of exposure across these three, meaningful time periods for the policy's implementation. I describe results estimating the policy's impact on each outcome: attendance rates, math and reading test score means, dropout rates, graduation rates, and promotion rates.

Attendance

As the policy implementation required significant shifts for students with disabilities, many of whom were transitioned into entirely new school buildings as well as into general education classrooms, there is a reasonable expectation that there would be equally significant disruption to student attendance—both for the moving students as well as their peers in classrooms with new student compositions. However, the results do not bear this out. While there are some statistically significant findings observed across all grades and within the school-level-specific results, the magnitude of the findings is so small as to have no substantive significance: both increases and decreases in attendance rates are within one percentage point in either direction. While the

[20] Following Sun and Abraham (2021), the weights underlying IW estimates are determined by sample shares of each cohort in each treatment period.

[21] Leads greater than 3 years prior to implementation and lags more than 9 years after the start of the policy implementation were binned. The time horizon is intentionally restricted to 9 years after the conclusion of implementation to allow for the observation of long-term impacts while also limiting the potential influence of additional factors that could occur much later and skew estimates.

estimates across school level models skew slightly more positive, particularly for students without disabilities, broadly speaking there were no substantive impacts on student attendance as a result of this policy. Figure 10 shows event study estimates for all students in grades 3 through 12; results for all other school levels and student groups are in Appendix E.

Math and Reading Test Scores

Contrary to expectations, estimates from all event study specifications showed no statistically significant changes in test scores for either student subgroup in reading or math. In other words: students with and without disabilities did no better or worse academically as a result of this policy implementation. This finding is reflected for math across all grades and student groups in Figure 11 and for reading in Figure 12. Results further disaggregated by elementary and middle school grades are in Appendix E. While there are other academic metrics that would shed insight into this policy's academic impacts (e.g., formative and summative subject assessments or test scores from other subject areas), based on the data available, it is clear that this policy of inclusion did no harm academically to any elementary or middle school students in the case study district.

Reardon and colleagues (2016) note that HOMOP transformations are limited in the cases of insufficient data and that imprecise estimates are possible when sample sizes fall below 50. As this impacts a key subgroup of interest in this study—students with disabilities, for whom within-school frequencies are naturally small—I ran this analysis again using only cell counts greater than or equal to 50. These results, presented in Appendix D, are consistent with initial findings; there were, again, no changes to either math or reading test score means as a result of the policy for either student group.

In addition to assessing changes in test scores for each individual subgroup as a result of the policy, it is also worth considering the impact of the policy on achievement gaps *between* the two groups—students with and without disabilities. In the previous event study results, achievement reflects recovered test score means transformed from percent-proficient among each subgroup through the homoskedastic ordered probit (HOMOP) method described in Section 3.1. These transformations make comparisons over time more reliable, but—as previously discussed—they are limited as they take place subgroup-by-subgroup, meaning the resulting test score means (and performance gaps between subgroups) are not directly interpretable from the previous results.

To speak to achievement gaps, following Ho (2009), I take the inverse normal of the original, state-reported percent-proficient for each subgroup and calculate the resulting achievement gaps on this scale of standard deviation units. Figure 13 presents the path of achievement gap trends between the two groups over time, again split by school level.[22] There are two notable takeaways from this visualization. First, during the policy implementation years, the achievement gap between students with and without disabilities generally declines across both reading and math. However, after 2009—the post-implementation years—these achievement gap trends reverse course and gaps between the two groups increase back to pre-policy levels or higher.

Dropout Rates

Absent consistent, standardized assessment data in high school subjects, alternative metrics are used to gauge policy impacts on students' academics at this level, including four-year adjusted cohort dropout rates.[23] During the implementation period, a two percentage-point increase in dropout rates is observed among students with disabilities ($p<.01$). However, this increase did not sustain for students with disabilities after the implementation concluded (see Figure 14). The policy also resulted in a one percentage point decrease in dropout rates among students without disabilities over the 5 to 9 years following implementation ($p<.01$).

It is possible that, for either behavioral or academic reasons, high school students with disabilities struggled more to acclimate during the initial years of inclusion than their peers without disabilities, resulting in their more frequent exit from schooling. However, this pattern did not sustain after implementation concluded, and students without disabilities saw a slight *decline* in dropout rates as a result of the policy—suggesting a longer-term positive effect for non-disabled students of time spent in classrooms with diverse learners. More research is needed to understand the implementation-period factors contributing to increased dropout rates among students with disabilities.

[22] Because this is a different method of transformation than that used to transform the raw data used in the event study analyses, the magnitude of gaps reported in this figure are not comparable to those in the event study results.

[23] This measure is defined as the number of dropouts (students who terminate formal education for any reason other than death and are not known to enroll in another school or state-approved program) divided by the adjusted student cohort (the number of first-time 9th graders, plus any students who transfer in, minus any who transfer out, emigrate or die during the four-year period).

Graduation Rates

The policy had no impact on graduation rates during the implementation period, but resulted in a 2.6 percentage point increase in the years following implementation ($p<.05$).[24] Figure 15 illustrates the steady increase in graduation rates over the nine-year, post-implementation period.

Figure 16 further highlights this long-term positive trend. While graduation rate data are unavailable for students with and without disabilities from the pre-policy and implementation time periods, disaggregated data on high school graduation rates are available from 2009 to 2019—the post-implementation years. As illustrated in the figure, high school graduation rate trends across the full population and both subgroups increase steadily in the subsequent decade following the conclusion of the policy implementation. This finding offers additional evidence that there were no detrimental academic impacts for students in the case study district in the long term, and further suggests that when looking beyond the lack of causal impacts on standardized test scores there may still be positive academic benefits for students with and without disabilities to education in inclusive classrooms. However, as with dropout rates, more information is needed to better understand the mechanisms of this relationship.

Promotion Rates

Results for student promotion rates from one grade to the next are consistent with results for both dropout and graduation rates. The policy had no impact on students' average likelihood of promotion to the next grade during the implementation period, but a positive, statistically significant impact in the longer-term. Students in 9th and 10th grade were 6.7 and 2.2 percentage points more likely, respectively, to be promoted to the next grade in the post-implementation years ($p<.01$). No impacts on student promotion were observed in the later high school grades. Results for 9th and 10th grade promotion rates are presented in Figure 17, and all other grades are in Appendix E. This is a final piece of evidence suggesting a positive policy impact on high school students' academics. While promotion rates are not a direct measure of academic proficiency, they

[24] The graduation rate measure reflects the percentage of students receiving a high school diploma, defined as the number of high school graduates divided by the sum of dropouts for grades 9 through 12 in consecutive years plus the number of high school graduates. This calculation, which is distinct from the calculation of an adjusted four-year cohort graduation rate, for example, accounts for dropout rates within the measure itself. Because of this, estimates of policy impacts on dropout rates cannot be directly compared to the estimates of impacts on high school graduation rates overall.

are by definition a measure of preparedness for the next grade level and are therefore a comprehensive measure, similar to graduation rates, of students' ability to succeed academically in inclusive learning environments.

5 Robustness Checks

A series of robustness checks were conducted to assess the previous results' sensitivity to the choice of comparison group as well as the potential confounding influences of: heterogeneous treatment effects, changes to the tested student population, changes to the population of students with disabilities, differential policy impacts across disability classifications, and the impacts of inclusion on student behavior.

Comparison Group Choice

Results against both alternative comparison groups—only other schools in rural districts in the state and a synthetically generated comparison group—are presented in Appendix A, and show some minor differences compared to the main results.

Comparing against all other rural schools in the state, attendance rates across all school levels remain substantively insignificant, with changes still inside one percentage point in either direction. Math and reading test scores also remain largely unchanged, though one coefficient—math scores for students without disabilities across all grades (3 through 8) during the implementation period—becomes statistically significant, reflecting a slight increase for these students' test scores of 0.09 standard deviations as a result of the policy ($p<.10$). Among high school students, dropout rates for students with disabilities are still shown to increase during the implementation period, but against the rural-only comparison group the increase in dropouts sustains in the post-implementation years, though at a lower rate than during the implementation years (1.3 versus 2.3 percentage points). Graduation and promotion rate estimates also shift, with results suggesting small decreases in both for all students during implementation, but no changes to either over the longer term.

Measuring policy effects against the synthetic comparison group, findings are even closer to those in the main results. Attendance rates are not substantively impacted at any school level, and math and reading test scores are again shown to be completely unaffected by the inclusion policy. At the high school level, dropout rates for students with disabilities increase during the

implementation period to a similar degree as the main results (3 percentage points versus 2.5), and again this does not sustain in the later years. Graduation and promotion rates again increase in the post-implementation period at rates similar to the primary findings.

Overall, though there are some changes to findings across these six outcomes, particularly against the all-rural comparison group, none are significant enough to alter the broad conclusions drawn about the impacts of this policy. By all accounts, there was no harm done to students in the case study district as a result of the policy.

Heterogeneous Treatment Effects

To address the issue of heterogeneous treatment effects in the main results, I disaggregate findings by school level and include year-by-school-type fixed effects. As an additional check, I estimated the main model using an interaction-weighted estimator, accounting for the weighted average of each cohort-average treatment-on-the-treated estimate (Sun & Abraham, 2021). Results from this estimation are in an Appendix B, and show small decreases in the magnitude of some coefficients but no changes to either substantive or statistical significance for any outcome. This confirms that the adjustments to the main model have accounted for the majority of this issue.

Testing

The years over which this policy took place overlap with significant federal changes to school accountability policies under No Child Left Behind (NCLB). NCLB required that all states test all students annually, and that results be disaggregated and reported for specific student subgroups including students with disabilities. This increased accountability mechanism likely drew new students into the tested-students sample, which would bias estimates of the inclusion policy's impact if students who were less likely to perform well on standardized assessments (e.g., students with more severe disabilities) were increasingly included in the sample.

Figure 18 shows the percentage of test takers over time for both the case study district and other districts in the state, for all students and students with disabilities. There is a notable increase in the number of test takers across all groups between 2003 and 2005, likely a result of NCLB accountability mechanisms slowly changing district and school behavior. To assess whether this descriptive increase in test takers is biasing results, I regressed participation rates as an outcome using the previous event study model. Results from this assessment are in Table 7 and show that

participation rates among students with disabilities did not change substantively during either the policy's implementation period or the 5 to 9 years after it concluded. While data are not available on the test-taking population by disability classification, this suggests results examining students' academic outcomes are not biased by changes in the number of students with disabilities participating in testing.

While the passage of NCLB overlapped with the beginning of the inclusion policy's implementation, the introduction of the Partnership for Assessment of Readiness for College and Career (PARCC) assessment overlapped with the end. In the 2014-15 school year, all schools in the case study state formally transitioned from using the state standardized assessment of the previous decade to the nationally-normed PARCC assessments as the statewide measure of students' academic performance. Pilot testing of the PARCC assessments began statewide in 2013-14, and teaching transitions to curriculum addressing the Common Core State Standards (standards aligned to the PARCC assessment, but not aligned to the previously used state assessment) began as early as 2012-13. While state test data are available from this period, student performance on the state assessment are not a reliable indicator of academic achievement and are less comparable to data from previous years. Results from the main model eliminating the years after 2012-13 are in Appendix C, and reflect some small increases in the magnitude of some coefficients, but no substantive changes to overall results. This implies that the influence of PARCC in the later years slightly, negatively biased the main results.

District Population Changes

An increase in the number of students with disabilities being removed from or moving out of the case study district in response to the policy would similarly bias results and mask the policy's true impact. Figure 19 displays the number of nonpublic placements in the case study district over time in two ways: as the raw number of nonpublic placements and as the number of nonpublic placements against the total number of students with disabilities in the district. The data in Panel A show a slight increase in the number of students with disabilities sent to private settings over the policy implementation period, followed by a steady decline back to pre-policy levels beginning in 2008. Panel B compares these numbers to the total number of students with disabilities within the district and demonstrates that while nonpublic placements increased over the policy

implementation period, the change was not meaningful with respect to the total number of students with disabilities who remained in the district's public schools under the new policy of inclusion.

Figure 20 illustrates the related issue of student mobility and shows that the case study district saw no significant change in overall mobility rates over the policy implementation period. Mobility rates measure the sum of student entrants and withdrawals over a total student population, and are therefore not a perfect indicator of the number of students exiting a district voluntarily. However, the consistency of the case study district's mobility rate means either the inclusion policy did not spur an increase in voluntary student exits or there was an increase in student withdrawals but it was masked by a comparable influx of new entrants each year—a mathematical improbability.

Disability Classifications

By 2006, roughly 90% of all students with disabilities in the case study district were placed in general education as their primary learning environment—a rate which sustained until 2019.[25] Student-level data from post-implementation years enable a more granular analysis of whether students across disability types were equally likely to be placed in inclusive classrooms, with some expected variation relative to their level of need. While this information is only available for the 2020-21 school year, given that the proportion of all students with disabilities in general education stayed consistent at high levels over a 13-year period and there is no evidence of students with disabilities disproportionately exiting the district in response to the policy, it is probable that the underlying composition of students in inclusive classrooms also did not substantively change over this time.

Figure 21 shows the distribution of disability classifications within inclusion settings as a proportion of all students with each classification in the population for the 2020-21 school year. Panel A suggests there is no systematic discrimination by disability classification in terms of likelihood of placement in general education settings. The majority of all students with disabilities, regardless of classification, are spending 80% or more of their school day in general education. Panel B reaffirms this conclusion, but also demonstrates that the likelihood of placement in inclusion varies across classifications, as expected. Smaller percentages of students with more severe disabilities (i.e., emotional disturbance, intellectual disabilities) are placed in inclusive

[25] 2019-20 is the most recent academic year for which educational environment data are available.

classrooms relative to peers with less severe disabilities. There are likely differential impacts of the inclusion policy among students with disabilities by classification; however, understanding these impacts requires more granular data beyond the scope of this study.

Student Behavior

Previous research on peer effects suggests that students with disabilities who also have behavioral issues have a largely deleterious effect on the academic outcomes of their general education classroom peers (Fletcher, 2010; Gottfried, 2014; Carrell, Hoekstra, & Kuka, 2016). The Civil Rights Data Collection (CRDC) has collected data on issues of behavior among students with disabilities (and other subgroups) and related punishments since 2000, though the format of this data changed substantively in 2009.[26] After 2009, data were disaggregated between students with and without disabilities whereas previous data were only available for students with disabilities. This information enables a descriptive assessment of behavioral issues among students in the case study district before, during, and after the policy was implemented.

Table 8 presents the average number of reported instances of discipline among students with and without disabilities in the case study district across all collected categories in each year. Due to the changes to the data collected over time, discipline data for students without disabilities are unavailable for the pre-policy years. For students with disabilities, instances of discipline increase steadily over time, whereas the reverse trend is true for their peers without disabilities. This suggests that in the period after the policy was fully implemented students with disabilities may have struggled with classroom behavior while their peers without disabilities steadily acclimated to the new policy. However, no causal conclusions can be drawn about the relationship between student behavior and the policy from these data alone. Additional analysis into the impact of inclusion on student behavior is worthwhile but beyond the scope of this study.

[26] In the years 2000, 2004, and 2006, CRDC data were collected indicating the number of instances of discipline associated with students with disabilities in each school in a given year for three categories: students who received corporal punishment, students who were suspended or expelled with educational services, and students who were suspended or expelled without educational services. Beginning in 2009, every two years this information was collected at more granular level indicating the number of students in each school in a given year who received: one or more out-of-school suspension, one or more in-school suspension, corporal punishment, a school-related arrest, an expulsion with or without educational services, a referral to law enforcement, a transfer to an alternative school, or an expulsion under a zero-tolerance policy. These data were also disaggregated not only by special education versus general education student groups, but also by students' primary race/ethnicity classification. As a result of this change to data collection, data from before 2009 are not easily comparable to the years prior. The smaller number of reporting categories in the earlier years naturally results in fewer reported instances.

6 Conclusion

There were no detrimental impacts for students with or without disabilities by any metric considered in this analysis. Put another way: the policy of inclusion was implemented and neither student group was made substantively worse off as a result. By the most common measure of academic performance, standardized test scores, there were no changes for any students in either reading or math as a result of general education classrooms becoming the default educational environment. This finding is contrary to expectations in related literature which suggest that when students with disabilities are moved into general education settings the academics of their peers without disabilities often suffers as a result. In this district, this neutral policy impact on academics implies the shift of the default student placement from excluded to included does not come at the cost of any student group's learning and removes that potential tradeoff from any calculus about whether such a policy should be implemented.

Ancillary measures of academic performance—attendance, dropout, graduation and promotion rates—reinforce the conclusion that this policy did no harm to students in the case study district. Attendance rates stayed largely consistent across groups over time, with some minor (less than one percentage point) fluctuations. While students with disabilities saw a slight increase in high school dropout rates (two percentage points) during the implementation period, this increase did not sustain beyond the four, initial years of the policy. Estimates for graduation and promotion rates further suggest positive, long-term policy impacts after the implementation period concluded. More data is needed to understand these latter impacts for specific student subgroups, but both metrics indicate a positive influence of the policy on average for all students.

Notably, there is more to understand about this policy's effectiveness than the available data can convey. One missing component is a better understanding of student behavior, as prior research suggests it is the challenging behaviors of students with more severe disabilities that are the mechanism underlying their peers' negatively affected academic performance. Additional information on student discipline referrals or indicators of socioemotional well-being would augment this analysis. Descriptively, data from the Civil Rights Data Collection suggest students without disabilities may have struggled less with the transition to inclusion than their peers with disabilities—a finding modestly supported by the disaggregated event study results for dropout and attendance rates—but this conclusion is not causal. Additionally, data limitations preclude

observation of the differential impacts of the policy experienced by students across disability classifications. This study does not address this important question, but future work should.

On net, when considering findings from this effectiveness analysis, it is clear that the case study district was successful in placing students with disabilities in inclusive classrooms and ensuring that, once there, students with disabilities' academic performance did not suffer as a result. Results suggest the policy went one step further than this, also ensuring that students *without* disabilities performed as well as or better than they did previously. A reasonable metric for judging this policy's overall success might be whether the policy did any harm, or beget any unintended consequences on the path to increasing inclusivity within the district. Findings demonstrate that no harm was done academically. Further, an inclusive community was created for students—exposing them to children different from themselves, and allowing students with disabilities the opportunity to feel truly included in their school communities. These intangible elements do not show up in test scores and other measures of academic performance, but the academic findings resulting from the quasi-experimental analyses in this study highlight that for a community that chooses to prioritize inclusion there may be no related, academic tradeoff.

References

Affleck, J. Q., Adams, A., Lowenbraum, S., & Madge, S. (1988). Integrated classroom versus resource model: Academic viability and effectiveness. *Exceptional Children, 54*(4): 339-349.

Angrist, J. D. & Pischke, J. (2015). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Athey, S., & Imbens, G. W. (2021). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*.

Ballis, B. & Heath, K. (2019). The long-run impacts of special education. EdWorkingPaper No. 19-151. Annenberg Institute at Brown University.

Brady, F. (2010). The influence of inclusion on language arts literacy and math achievement of non-disabled middle school students. Doctoral Dissertation, Seton Hall University.

Brewton, S. (2005). The effects of inclusion on mathematics achievement of general education students in middle school. Doctoral Dissertation, Seton Hall University.

Callaway, B., & Sant'Anna, P. H. C. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Carrell, S. E., Hoekstra, M., & Kuka, E. (2016). The long-run effects of disruptive peers. Working Paper No. 22042. National Bureau of Economic Research.

Crockett, J. B. (2020). "Inclusion as an idea and its justification in law." In Kauffman, J. M. (Ed.), *On educational inclusion: Meanings, history, issues and international perspectives*. New York, NY: Routledge.

Daniel, L. & King, D. (1997). Impact of inclusion education on academic achievement, student behavior and self-esteem, and parental attitudes. *Journal of Educational Research, 91*, 67-80.

De Chaisemartin, C., & D'Haultfœuille, X. (2018). Two-way fixed effects estimators with heterogenous treatment effects. *American Economic Review, 110*(9): 2964-2996.

Dessemontet, R. S., Bless, G. & Morin, D. (2012). Effects of inclusion on the academic achievement and adaptive behaviour of children with intellectual disabilities. *Journal of Intellectual Disability Research, 56*(6): 579-587.

Dorn, S., Fuchs, D., & Fuchs, L. S. (1996). A historical perspective on special education reform. *Theory into Practice, 35*(1): 12-19.

Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., DiSalvo, R., & Ho, A. D. (2017). Stanford Education Data Archive. Technical Documentation, Version 2.0.

Fletcher, D. (2010). Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school. *Journal of Policy Analysis and Management, 29*(1): 69-83.

Gilmour, A. F. & Henry, G. T. (2018). Who are the classmates of students with disabilities in elementary mathematics classrooms? *Remedial and Special Education, 41*(1): 18-27.

Giordano, G. (2007). *American special education: A history of early political advocacy.* New York, NY: Peter Lang.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Gottfried, M. A. (2014). Classmates with disabilities and students noncognitive outcomes. *Educational Evaluation and Policy Analysis, 36*(1): 20-43.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? *The Review of Economics and Statistics, 84*(4): 584-599.

Helge, D. I. (1981). Problems in implementing comprehensive special education programing in rural areas. *Exceptional Children, 47*(7): 514-520.

Ho, A.D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6): 351-360.

Ho, A. D. (2009). Nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*(2): 201-228.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 et seq. (2004).

Jenkins, J. R., Jewell, M., Leicester, N., Jenkins, L., & Troutner, N. M. (1991). Developing a school building model for educating students with handicaps and at-risk students in general education classrooms. *Journal of Learning Disabilities, 24*(5): 311-320.

Jones, N. & Winters, M. A. (2020). Are two teachers better than one? The effect of co-teaching on students with and without disabilities. Wheelock Educational Policy Center, Working Paper 2020-1. Boston University.

Kurth, J. A., Morningstar, M. E. & Kozleski, E. B. (2015). The persistence of highly restrictive special education placements for students with low-incidence disabilities. *Research and Practice for Persons with Severe Disabilities, 39*(3): 227-239.

McDonnell, J., Thorson, N., Disher, S., Mathot-Buckner, C., Mendel, J., & Ray, L. (2003). The achievement of students with developmental disabilities and their peers without disabilities in inclusive settings: An exploratory study. *Education and Treatment of Children, 26*(3): 224-236.

McLeskey, J., Rosenberg, M. S. & Westling, D. L. (2018). *Inclusion: Effective Practices for all Students.* 3rd ed. Columbus, OH: Pearson.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education, 43*(4): 236-254.

National Center for Educational Statistics. (2019a). *The condition of education: Children and youth with disabilities.*

National Center for Educational Statistics. (2019b). Education Demographic and Geographic Estimates (EDGE), ACS-ED District Demographic Dashboard.

Peltier, G. L. (1997). The effect of inclusion of non-disabled children: A review of the research. *Contemporary Education, 68*(4): 234-239.

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2016). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. Stanford Center for Education Policy Analysis. Working Paper No. 16-02.

Robinson, C. M. (2012). The influence of inclusion on the academic performance of general education students on the New Jersey assessment of skills and knowledge in grades 6, 7, and 8. Doctoral Dissertation, Seton Hall University.

Roth, J. (2019). Pre-test with caution: Event study estimates after testing for parallel trends. Unpublished Manuscript. Department of Economics, Harvard University.

Salend, S. J. & Duhaney, L. G. (1999). The impact of inclusion on students with and without disabilities and their educators. *Remedial & Special Education, 20*(2): 114-127.

Sanger, C. S. (2020). "Inclusive pedagogy and universal design approaches for diverse learning environments." In Sanger, C. S. & Gleason, N. W. (Eds.), *Diversity and inclusion in global higher education: Lessons from across Asia.* Singapore: Palgrave Macmillan.

Sant'Anna, P. H. C., & Zhao, J. B. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics, 2019*(1): 101-122.

Schifter, L. A. (2015). Using survival analysis to understand graduation of students with disabilities. *Exceptional Children, 82*(4).

Schifter, L. A. & Hehir, T. (2018). "The Better Question: How Can We Improve Inclusive Education?" *Education Next*.

Schulte, A. & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children, 81*(3): 370-387.

Schwartz, A. E., Hopkins, B. G. & Stiefel, L. (2019). The effects of special education on the academic performance of students with learning disabilities. EdWorkingPaper No. 19-86. Annenberg Institute at Brown University.

Sharpe, M. N., York, J. L., & Knight, J. (1994). Effects of inclusion on the academic performance of classmates without disabilities: A preliminary study. *Remedial and Special Education, 15*(5): 281-287.

Smith, P. (2007). Have we made any progress? Including students with intellectual disabilities in regular education classrooms. *Intellectual and Developmental Disabilities, 45*(5): 297-309.

Sun, L. (2020). EVENTSTUDYWEIGHTS: Stata module to estimate the implied weights on the cohort-specific average treatment effects on the treated (CATTs) (event study specifications). In: Statistical Software Components. Boston College Department of Economics.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2021): 175-199.

The School Superintendent's Association (AASA) (2017). *Leveling the Playing Field for Rural Students*.

Tremblay, P. (2012). Comparative outcomes of two instructional models for students with learning disabilities: inclusion with co-teaching and solo taught special education. *Journal of Research in Special Educational Needs, 13*(4): 251-258.

Winzer, M. A. (2012). *History of special education: From isolation to integration.* Washington, DC: Gallaudet University Press.

Workforce Innovation and Opportunity Act, 29 U.S.C., Chapter 32 § 3101 et seq. (2014).

Workforce Investment Act, 29 U.S.C. § 2871 et seq. (2014).

Table 1: Case Study District Demographics

	Mean
Race/Ethnicity	
% AIAN	1.3
% Asian	1.7
% Black	10.4
% NHPI	0.4
% White	85.2
% Two+ races	6.4
% Hispanic	7.5
Gender	
% Male	52.3
% Female	47.7
Special Education	
% Students with IEPs	16.3
% Students receiving 504 accommodations	3.8
Other Demographics	
% ELL	2.5
% FRPL	49.3
% Title I	23.6
% Homeless	4.1
% Foster	0.5
% Gifted	6.7

Notes. This table presents demographic averages for the case study district based on student-level records from the district for the 2019-20 school year.

Table 2: Summary Statistics—Case Study District and Comparison Groups

	Case Study District	Comparison Group A *(Untreated schools in state)*	Comparison Group B *(Untreated schools in rural districts in state)*	Comparison Group C *(Synthetic comparison group)*
Avg. District population	99,069	555,557	61,993	571,690
Avg. School enrollment	566	646	545	662
% SPED	14.2	12.4	12.4	12.4
% FRPL	24.2	35.9	32.6	33.6
% AIAN	0.31	0.39	0.26	0.42
% Asian	0.68	4.36	0.99	4.99
% Hispanic	2.12	5.87	1.55	6.71
% Black	6.96	31.1	18.9	33.2
% White	85.2	45.7	72.2	49.9
Median Household Income	$65,079	$70,424	$65,441	$73,176
% Poverty	7.9	8.7	9.3	7.9
Number of districts	1	21	6	21
Number of schools	30	1,723	117	1,087

Notes. This table presents summary statistics for the case study district and three potential comparison groups, using data from the NCES Elementary and Secondary Information System (ELSI) from the 2002-03 school year (the year in which policy implementation began in the case study district). Data reflect averages across all schools in each group.

Table 3: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5-9 years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.409**	-0.0670	0.429**	0.610***	0.0768	0.640***
	(0.179)	(0.355)	(0.191)	(0.196)	(0.314)	(0.211)
Pre-treatment mean	92.8	91.1	93	92.8	91.1	93
Number of schools	1,606	1,579	1,562	1,606	1,579	1,562
Math Test Score Means	0.0432	0.0223	0.0867	0.0327	0.0209	0.0638
	(0.0429)	(0.0858)	(0.0536)	(0.0445)	(0.0848)	(0.0584)
Pre-treatment mean	0.020	-0.023	0.029	0.020	-0.023	0.029
Number of schools	1,057	1,043	1,050	1,057	1,043	1,050
Reading Test Score Means	-0.0117	0.0530	0.0233	-0.0139	0.0228	0.0168
	(0.0331)	(0.0669)	(0.0376)	(0.0416)	(0.0748)	(0.0461)
Pre-treatment mean	-0.008	-0.069	-0.014	-0.008	-0.069	-0.014
Number of schools	1,057	1,055	1,050	1,057	1,055	1,050

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across grades 3 through 12 for attendance and 3 through 8 for math and reading test scores. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 4: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.0152	-0.0658	0.0728	-0.100	-0.624***	-0.115**
	(0.0570)	(0.157)	(0.0594)	(0.0746)	(0.187)	(0.0575)
Pre-treatment mean	94.5	94.5	95	94.5	94.5	95
Number of schools	1,028	1,505	1,503	1,028	1,505	1,503
Math Test Score Means	0.0657	0.0117	0.120	0.0511	0.00953	0.0917
	(0.0584)	(0.126)	(0.0745)	(0.0584)	(0.118)	(0.0812)
Pre-treatment mean	0.01	-0.041	-0.023	0.01	-0.041	-0.023
Number of schools	799	797	797	799	797	797
Reading Test Score Means	-0.0135	0.0676	0.0311	-0.0149	0.0428	0.0333
	(0.0477)	(0.103)	(0.0556)	(0.0591)	(0.111)	(0.0682)
Pre-treatment mean	-0.030	-0.119	-0.035	-0.030	-0.119	-0.035
Number of schools	799	797	797	799	797	797

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table 5: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.000920	-0.175	-0.0242	0.368**	0.290	0.353
	(0.134)	(0.314)	(0.153)	(0.170)	(0.456)	(0.245)
Pre-treatment mean	93.6	90.4	94	93.6	90.4	94
Number of schools	493	480	457	493	480	457
Math Test Score Means	0.00590	0.0406	0.0321	0.00184	0.0455	0.0181
	(0.0588)	(0.0951)	(0.0737)	(0.0686)	(0.123)	(0.0759)
Pre-treatment mean	0.049	0.026	0.044	0.049	0.026	0.044
Number of schools	451	374	446	451	374	446
Reading Test Score Means	-0.0129	0.0342	0.00752	-0.0197	-0.00277	0.00224
	(0.0421)	(0.0490)	(0.0483)	(0.0504)	(0.0693)	(0.0473)
Pre-treatment mean	0.053	0.072	0.047	0.053	0.072	0.047
Number of schools	451	447	446	451	447	446

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 6: Event Study Results (High)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.115	-0.993***	0.189	1.059***	0.0959	1.109***
	(0.255)	(0.337)	(0.253)	(0.322)	(0.377)	(0.329)
Pre-treatment mean	90.6	88.9	90.9	90.6	88.9	90.9
Number of schools	353	332	316	353	332	316
Dropout Rate	0.224	2.505***	-0.383	-0.551	0.648	-0.826***
	(0.389)	(0.460)	(0.332)	(0.427)	(0.519)	(0.301)
Pre-treatment mean	4.2	9.9	3.3	4.2	9.9	3.3
Number of schools	369	347	341	369	347	341
Graduation Rate	-1.755	-	-	2.631**	-	
	(1.301)			(1.266)		-
Pre-treatment mean	81.3			81.3		
Number of schools	325			325		
Promotion Rate *(All Grades)*	0.451	-	-	2.945***	-	
	(0.920)			(1.020)		-
Pre-treatment mean	91.3			91.3		
Number of schools	379			379		
9th Grade	1.650*	-	-	6.738***	-	
	(0.959)			(1.298)		-
Pre-treatment mean	87.7			87.7		
Number of schools	352			352		
10th Grade	-1.243	-	-	2.195***	-	
	(0.973)			(0.838)		-
Pre-treatment mean	89.3			89.3		
Number of schools	336			336		
11th Grade	-0.170	-	-	0.983	-	
	(0.804)			(0.675)		-
Pre-treatment mean	93.4			93.4		
Number of schools	340			340		
12th Grade	-0.249	-	-	0.0954	-	
	(0.607)			(0.751)		-
Pre-treatment mean	94.6			94.6		
Number of schools	332			332		

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 9 through 12. Outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5

to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table 7: Testing Participation Rates

	Implementation Period (0-4 years)		Post-Implementation (5+ years)	
	All Students	*SWD*	*All Students*	*SWD*
Math Participation Rate	-1.223	0.0485	-1.187	0.658
	(1.193)	(2.943)	(1.196)	(2.811)
Pre-treatment mean participation rate	44.3	46.6	44.3	46.6
Reading Participation Rate	-1.177	0.178	-1.203	1.081
	(1.227)	(2.542)	(1.236)	(2.599)
Pre-treatment mean participation rate	44.3	45.7	44.3	45.7

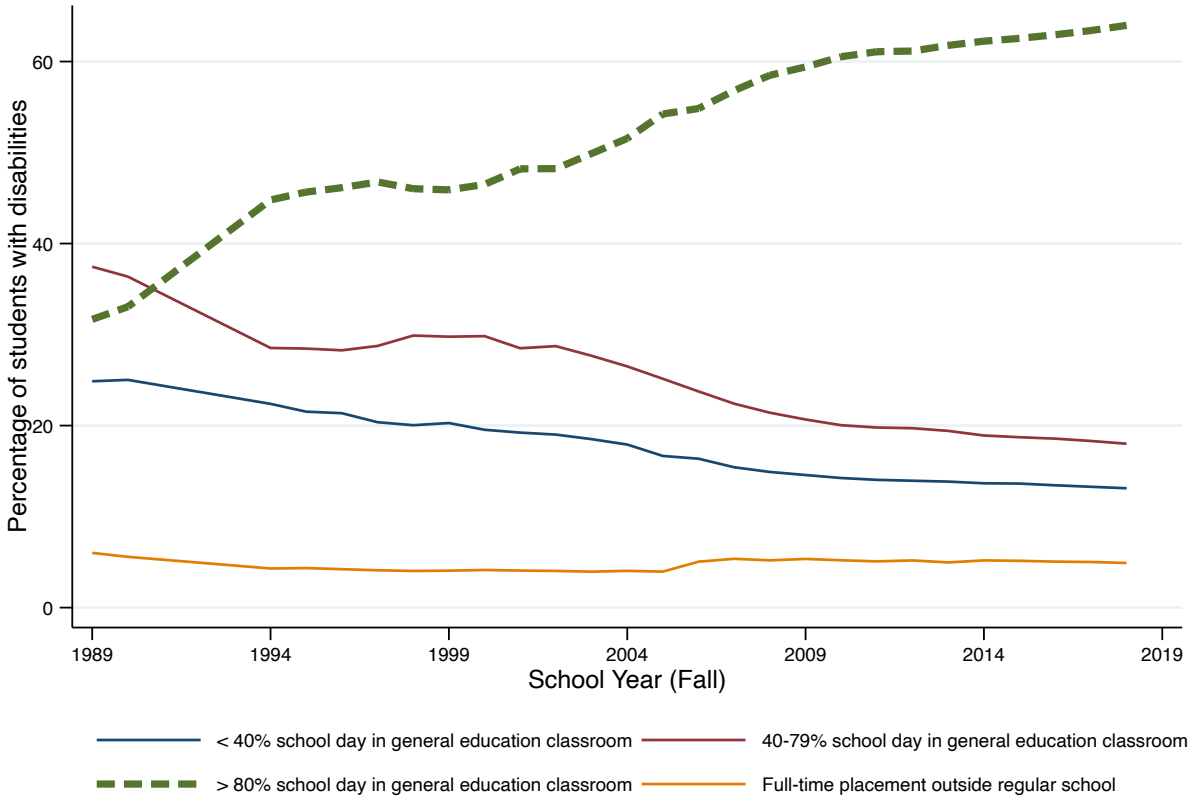
Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on standardized test participation rates in reading and math for all students and students with disabilities (SWD) across grades 3 through 8 in the case study district. Participation rate data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table 8: Annual Reported Instances of Discipline in Case Study District

	Pre-Policy	Policy Implementation			Post-Policy Implementation			
	2000	*2004*	*2006*	*2009*	*2011*	*2013*	*2015*	*2017*
Students with disabilities	<1	1.1	1.7	10	13.2	15.2	15.5	19.6
Students without disabilities	*N/A*	*N/A*	*N/A*	108	55.2	50.1	33.4	32.8

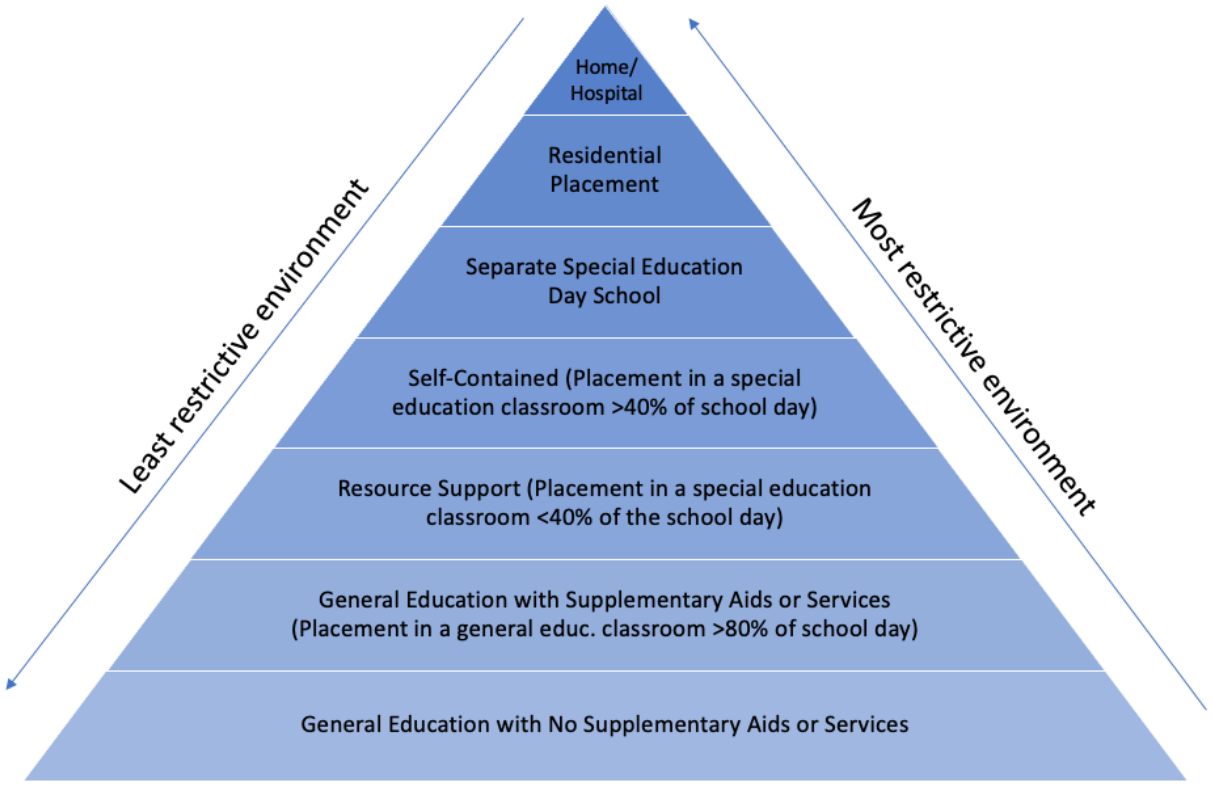
Notes. This table reports the number of reported instances of discipline among students with and without disabilities in the case study district between 2000 and 2017, based on school-level data from the Civil Rights Data Collection (CRDC). Data were not disaggregated across these two subgroups until 2009. Between 2000 and 2006 instances of discipline include reports of students with disabilities who: received corporal punishment, were suspended or expelled with educational services, or were suspended or expelled without educational services. After 2009, data have been collected every two years and include reports of students with and without disabilities who received: one or more out-of-school suspension, one or more in-school suspension, corporal punishment, a school-related arrest, an expulsion with or without educational services, a referral to law enforcement, a transfer to an alternative school, or an expulsion under a zero-tolerance policy.

Figure 1: Nationwide Prevalence of Inclusive Education



Notes. This figure illustrates national educational placement trends among students with disabilities ages 6-21 using historical data from the NCES Digest of Education Statistics. Data are grouped into four categories. The first three categories reflect the percentage of the school day students spend in general education settings across the three federal reporting categories—more than 80%, 40-79%, or less than 40%. Prior to 2008, these three reporting categories reflect time spent in general education at more than 60%, 21-60%, and less than 21%, respectively. The final category groups all placements where students spend 100% of the school day in a non-public-school setting (e.g., separate school, separate residential facility, private school, homebound, hospital, or correctional facility).

Figure 2: Continuum of Special Education Service Provision



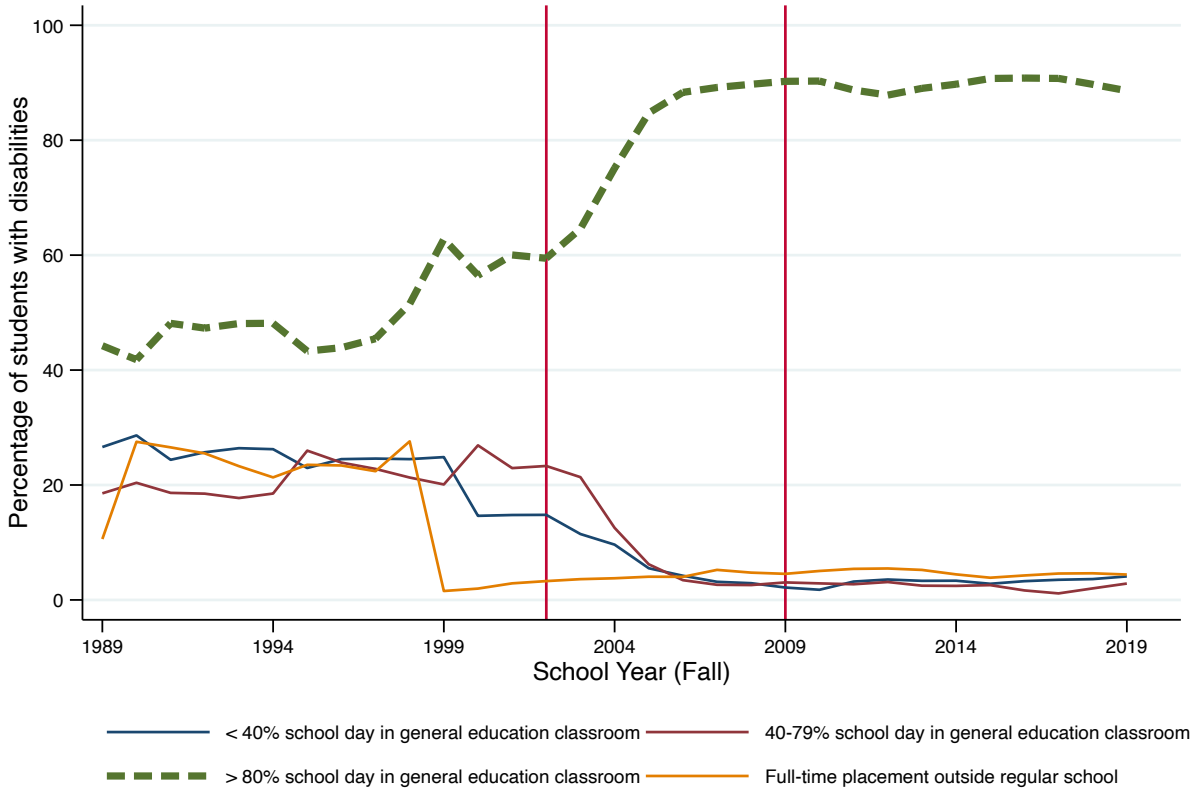
Notes. This figure illustrates the continuum of potential educational placements for students with disabilities from most restrictive to least restrictive, highlighting that the fewest number of students with disabilities should be placed into the most restrictive settings. The second tier from the bottom of this pyramid—general education with supplementary aids or services—reflects an "inclusive" educational placement.

Figure 3: Implementation Plan for Case Study District Schools

	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10
Year 1: Planning and training	A Elementary B Elementary C Elementary C Middle	D Elementary E Elementary F Elementary G Elementary	H Elementary I Elementary J Elementary K Elementary L Elementary M Elementary E Middle G Middle	Q Middle O Middle N Elementary O Elementary P Elementary	R Middle S Middle R High S High E High C High G High			
Year 2: Implementation and continued training support		A Elementary B Elementary C Elementary C Middle	D Elementary E Elementary F Elementary G Elementary	H Elementary I Elementary J Elementary K Elementary L Elementary M Elementary E Middle G Middle	Q Middle O Middle N Elementary O Elementary P Elementary	R Middle S Middle R High S High E High C High G High		
Year 3: Implementation and follow up			A Elementary B Elementary C Elementary C Middle	D Elementary E Elementary F Elementary G Elementary	H Elementary I Elementary J Elementary K Elementary L Elementary M Elementary E Middle G Middle	Q Middle O Middle N Elementary O Elementary P Elementary	R Middle S Middle R High S High E High C High G High	
Year 4: Continued training and support				A Elementary B Elementary C Elementary C Middle	D Elementary E Elementary F Elementary G Elementary	H Elementary I Elementary J Elementary K Elementary L Elementary M Elementary E Middle G Middle	Q Middle O Middle N Elementary O Elementary P Elementary	R Middle S Middle R High S High E High C High G High

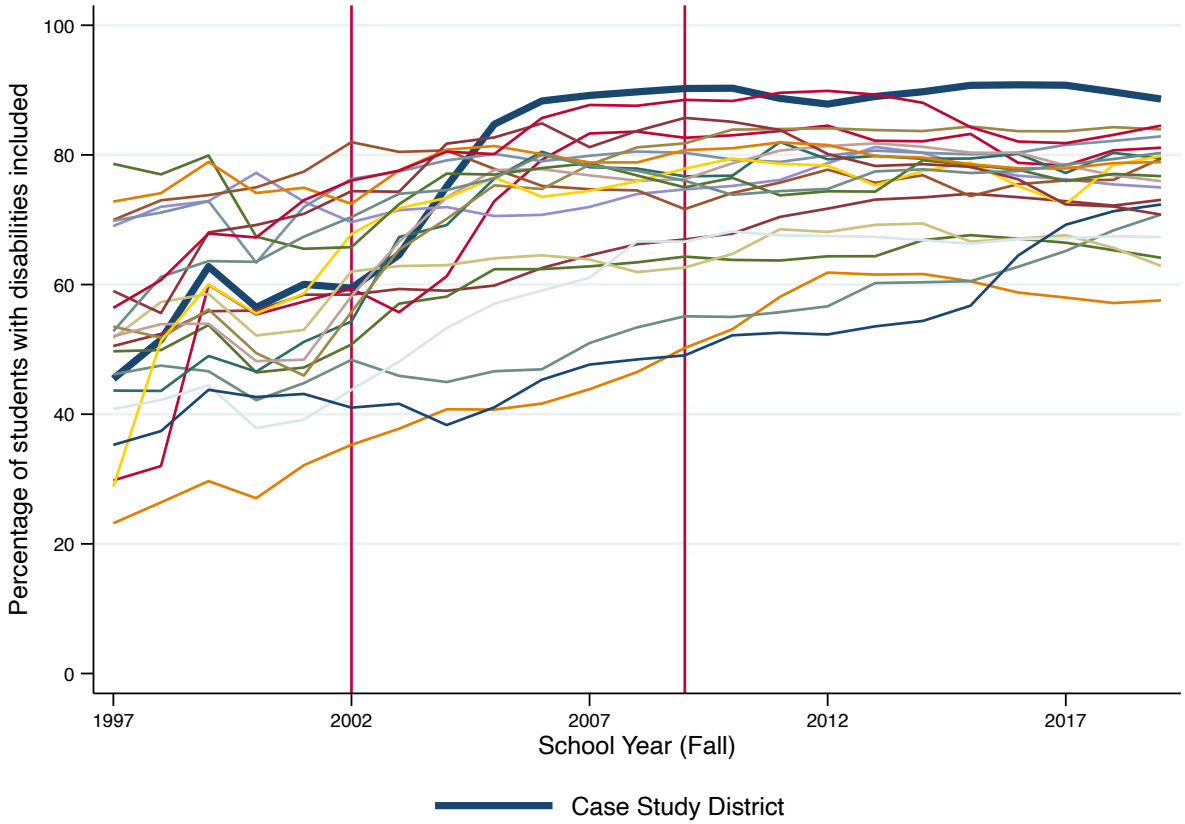
Notes. This chart reflects an artifact from the case study district outlining the plan for policy implementation across all district schools, but with the names of the schools redacted to preserve district anonymity. Each school in the district followed a four-year implementation arc and all district schools completed the policy implementation between the 2002-03 and 2009-10 school years.

Figure 4: Prevalence of Inclusion in Case Study District



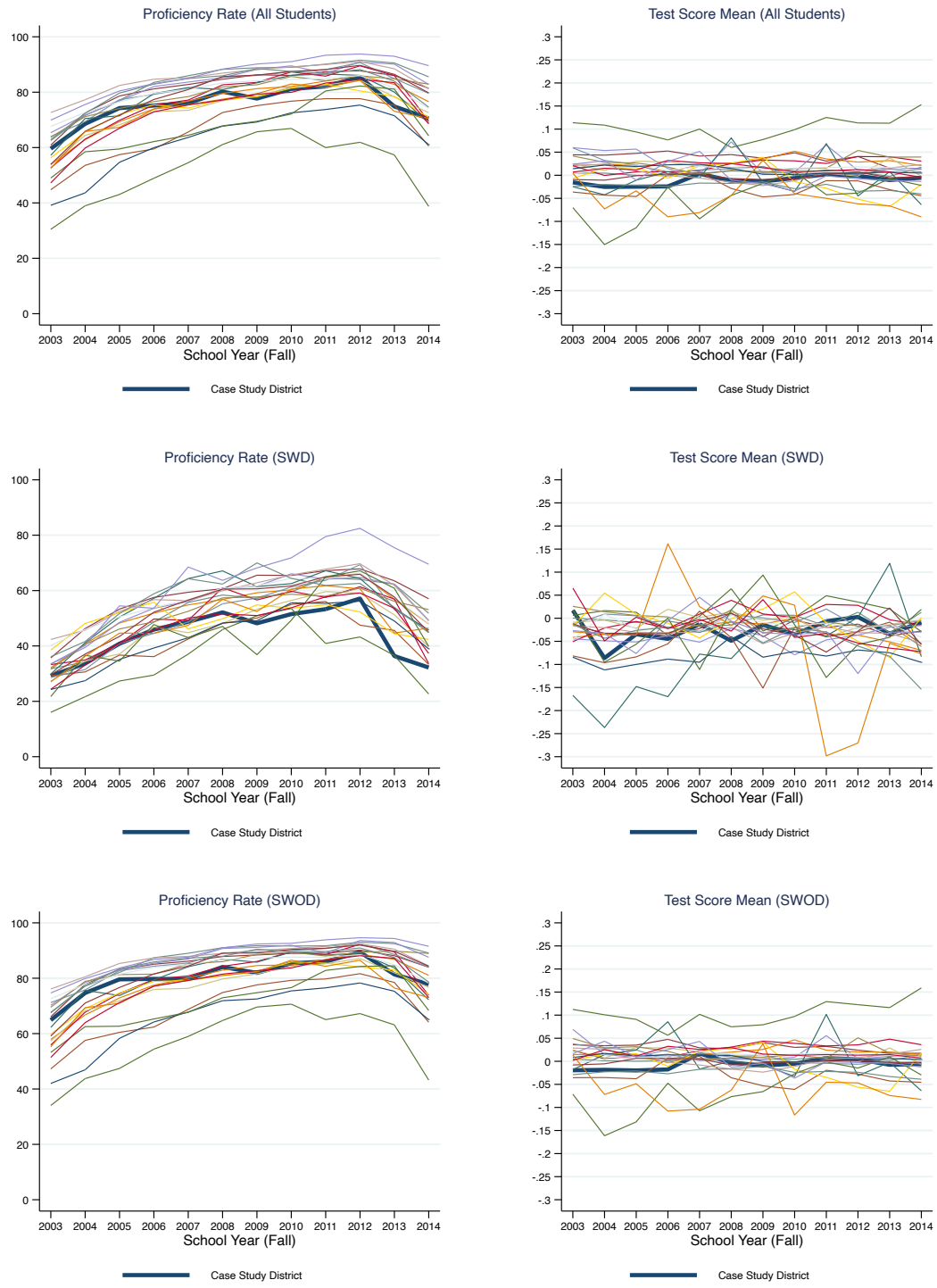
Notes. This figure illustrates educational placement trends among students with disabilities ages 6-21 using historical data from the case study state department of education. Data are grouped into four categories. The first three categories reflect the percentage of the school day students spend in general education settings across the three federal reporting categories—more than 80%, 40-79%, or less than 40%. Prior to 2008, these three reporting categories reflect time spent in general education at more than 60%, 21-60%, and less than 21%, respectively. The final category groups all placements where students spend 100% of the school day in a non-public-school setting (e.g., separate school, separate residential facility, private school, homebound, hospital, or correctional facility). Vertical lines at 2002 and 2009 mark the beginning and end of the policy implementation period.

Figure 5: Prevalence of Inclusion in All Districts in Case Study State



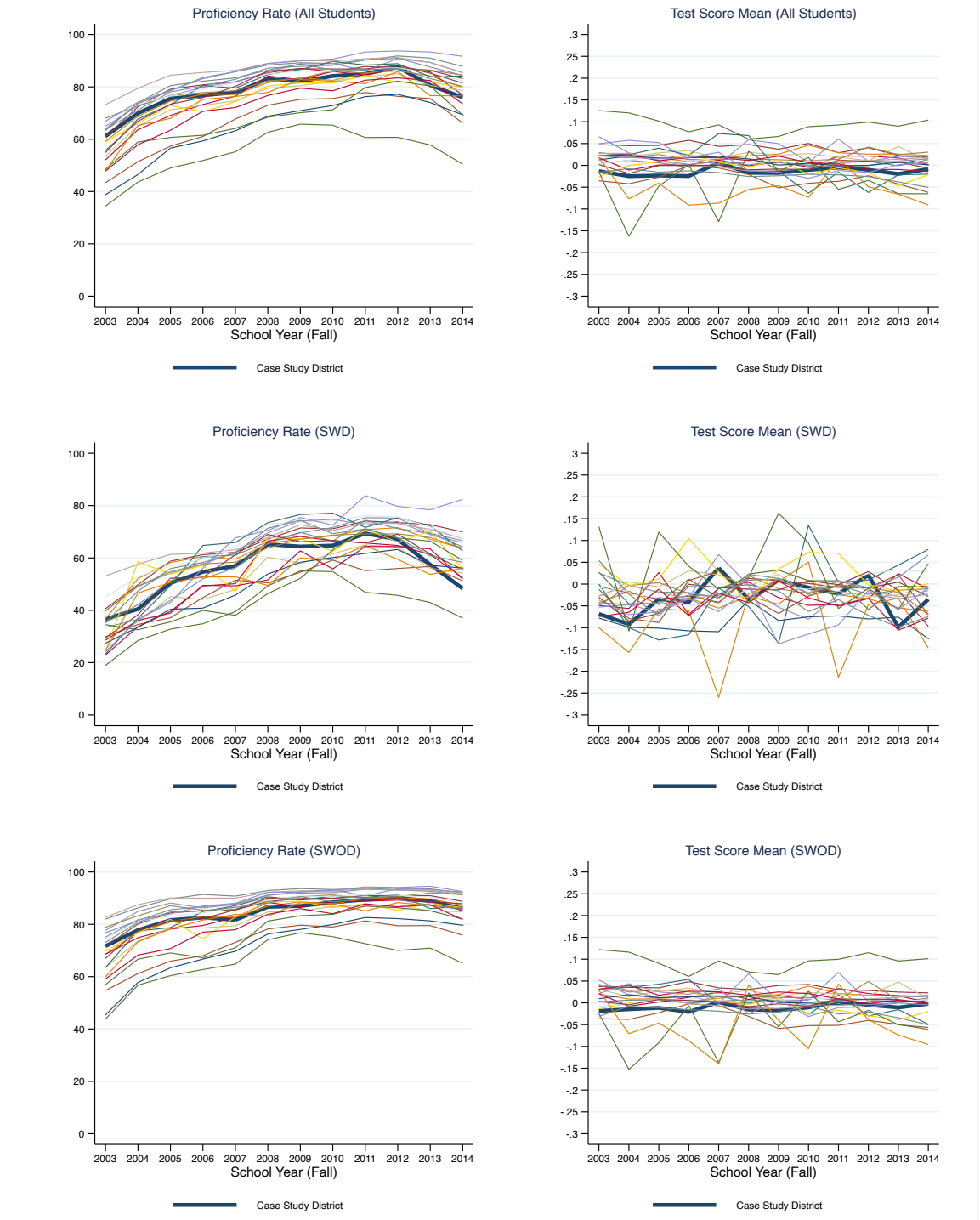
Notes. This figure illustrates the percentage of students with disabilities ages 6-21 whose primary educational placement is general education for more than 80% of the school day (i.e., an inclusive placement) for all school districts in the case study state. Data are from the case study state department of education. Vertical lines at 2002 and 2009 mark the beginning and end of the policy implementation period.

Figure 6: Trends in Math Performance—Proficiency Rates versus Recovered Means



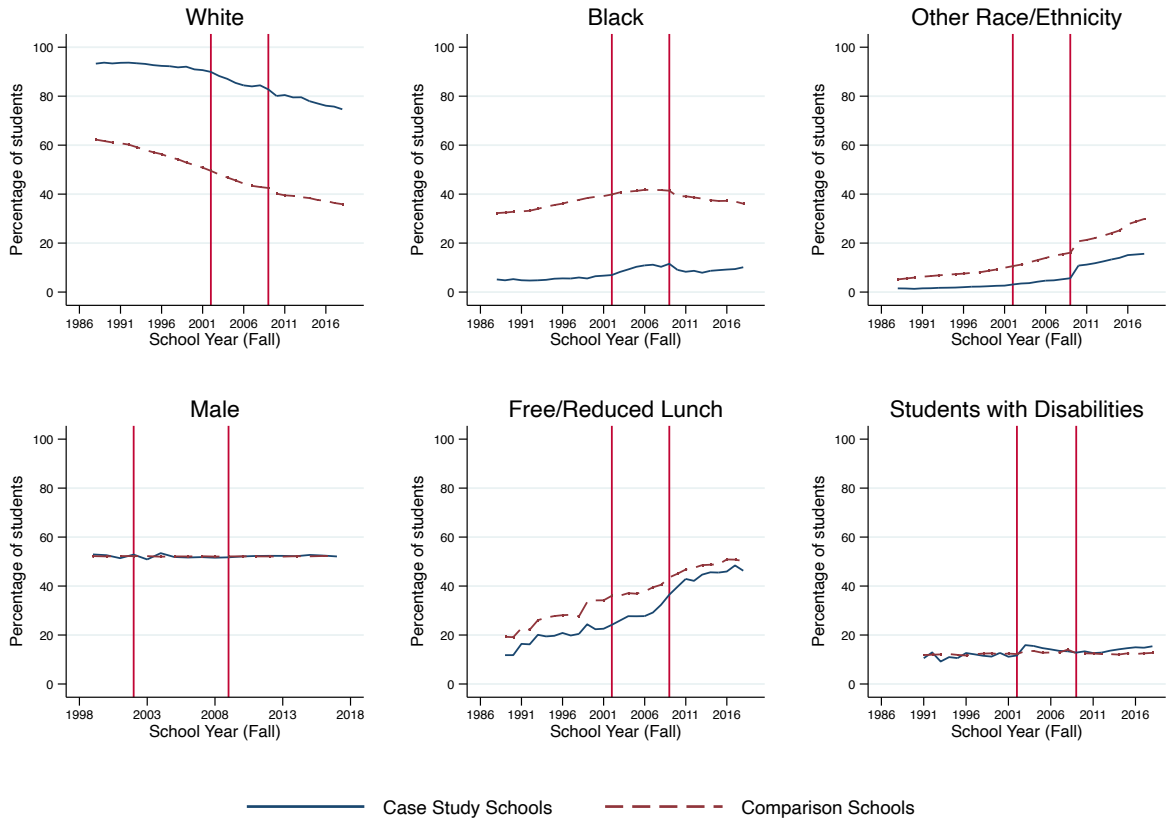
Notes. This figure illustrates student academic performance on standardized math assessments for all school districts in the case study state from 2003 to 2014 represented two ways: (1) as average proficiency rates by subgroup, as reported by the case study state, and (2) as recovered test score means based on homoskedastic ordered probit transformations of the data. Performance trends are broken down by three subgroups—all students, students with disabilities, and students without disabilities—for students in grades 3 through 8.

Figure 7: Trends in Reading Performance—Proficiency Rates versus Recovered Means



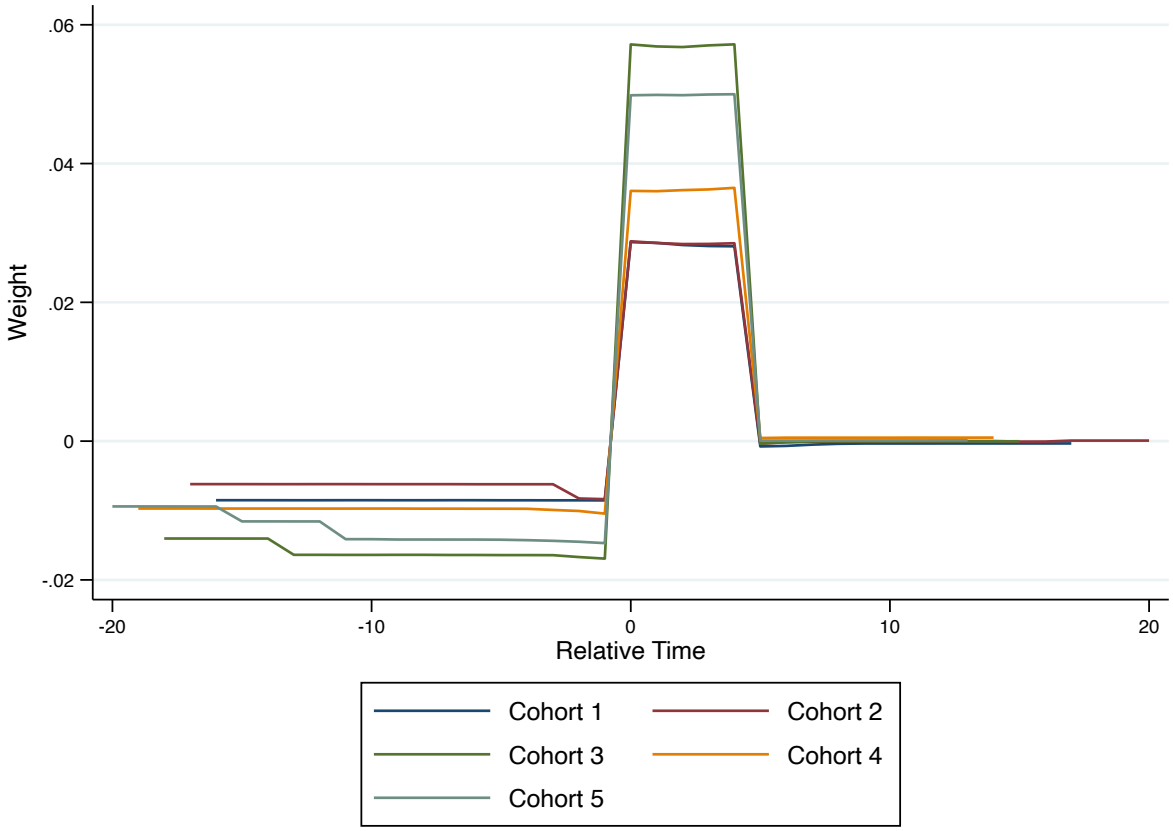
Notes. This figure illustrates student academic performance on standardized reading assessments for all school districts in the case study state from 2003 to 2014 represented two ways: (1) as average proficiency rates by subgroup, as reported by the case study state, and (2) as recovered test score means based on homoskedastic ordered probit transformations of the data. Performance trends are broken down by three subgroups—all students, students with disabilities, and students without disabilities—for students in grades 3 through 8.

Figure 8: Parallel Trends in Enrollment Demographics



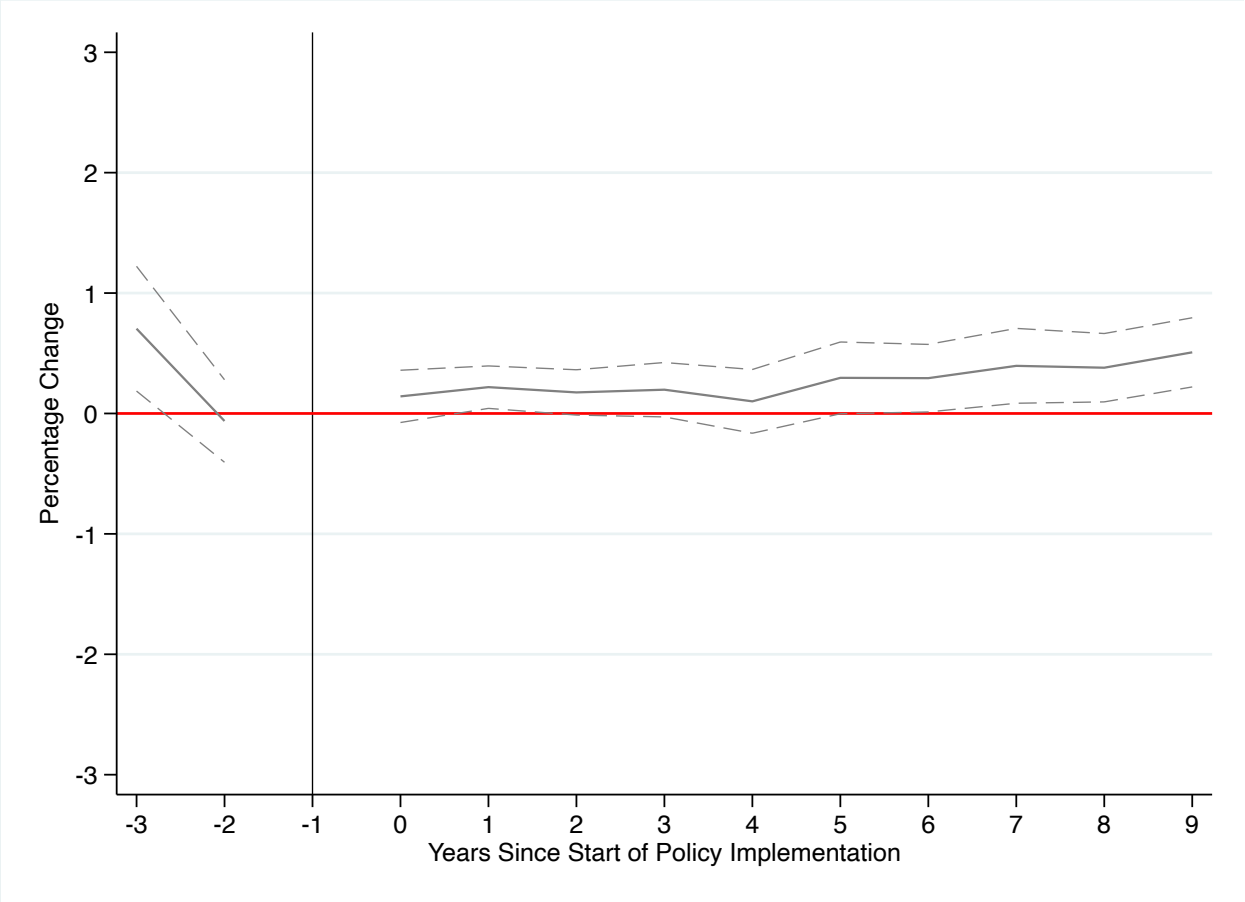
Notes. This figure presents demographic averages for the case study district and comparison group A—all other untreated schools in the state—based on student-level records from the case study state for the 2019-20 school year.

Figure 9: Estimated CATT Weights Underlying Event Study Specification



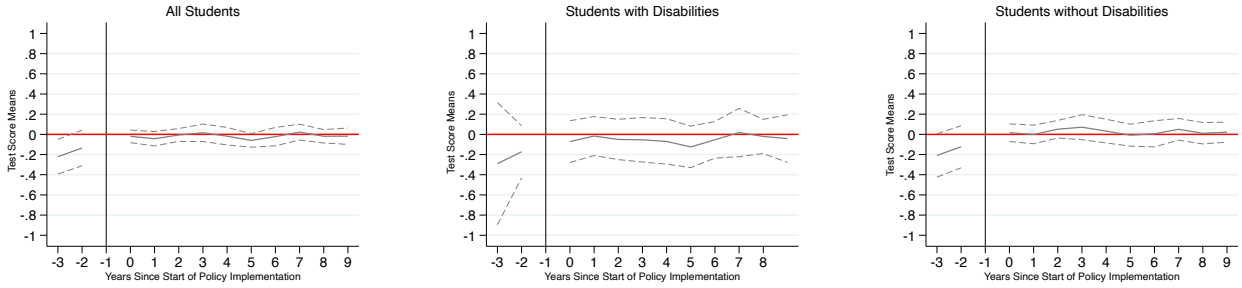
Notes. This figure illustrates the cohort-specific average treatment on the treated (CATT) weights underlying the TWFE regressions within the main event study specification in this study. CATTs are shown for each of the five treated cohorts within the case study district during the policy implementation period (2002-2009). Following Sun and Abraham (2021), these weights are calculated through an auxiliary regression depending on the distribution of the cohorts and indicators of relative time, using the `eventstudyweights` command in Stata (Sun, 2020).

Figure 10: Attendance Rates—All Students



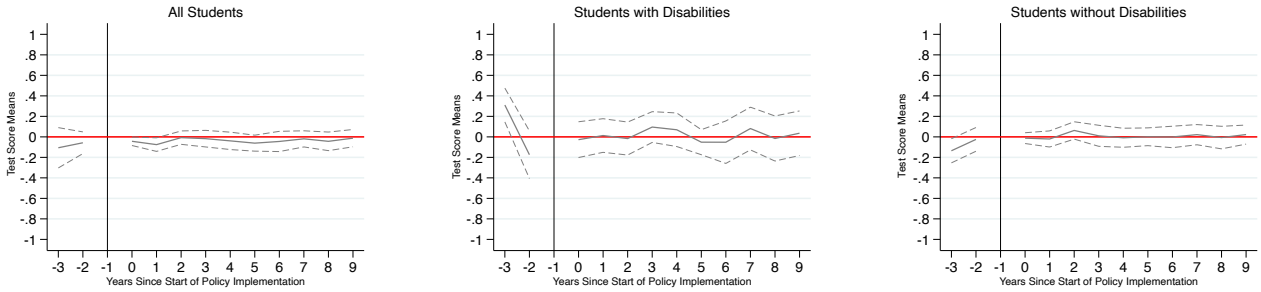
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on attendance rates for all students in grades 3 through 12 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 11: Math Test Score Means



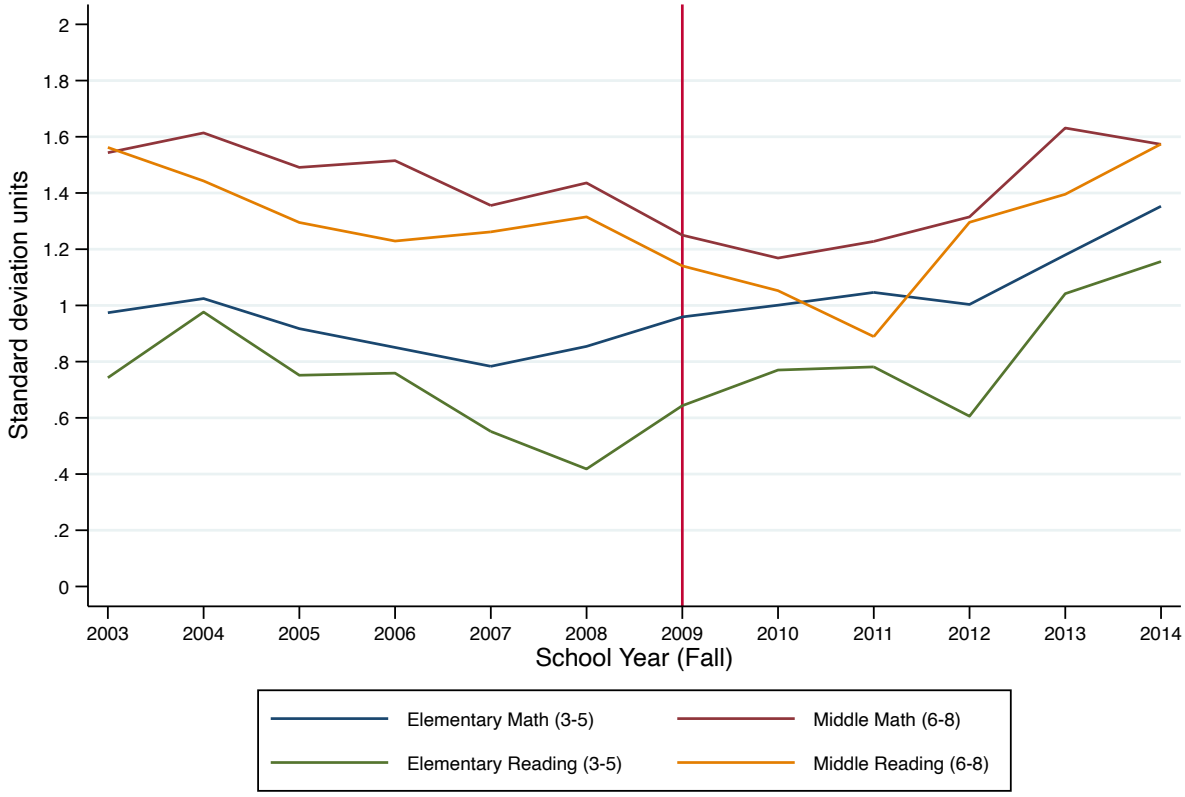
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on math test score means for all students in grades 3 through 8 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 12: Reading Test Score Means



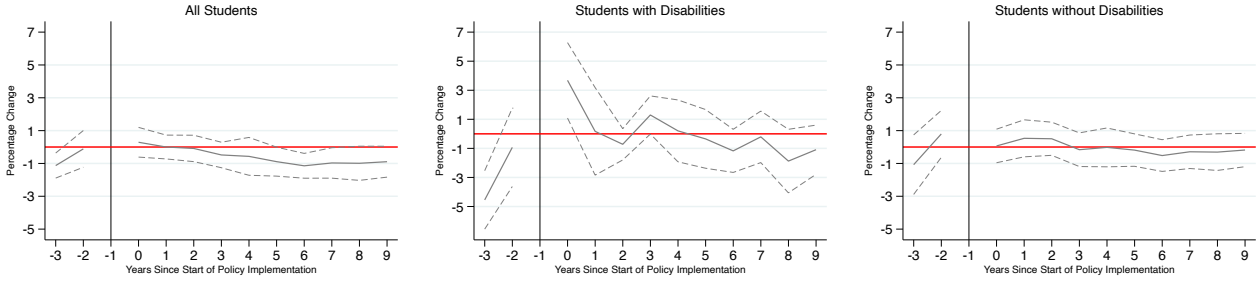
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on reading test score means for all students in grades 3 through 8 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 13: Student Achievement Gap Trends



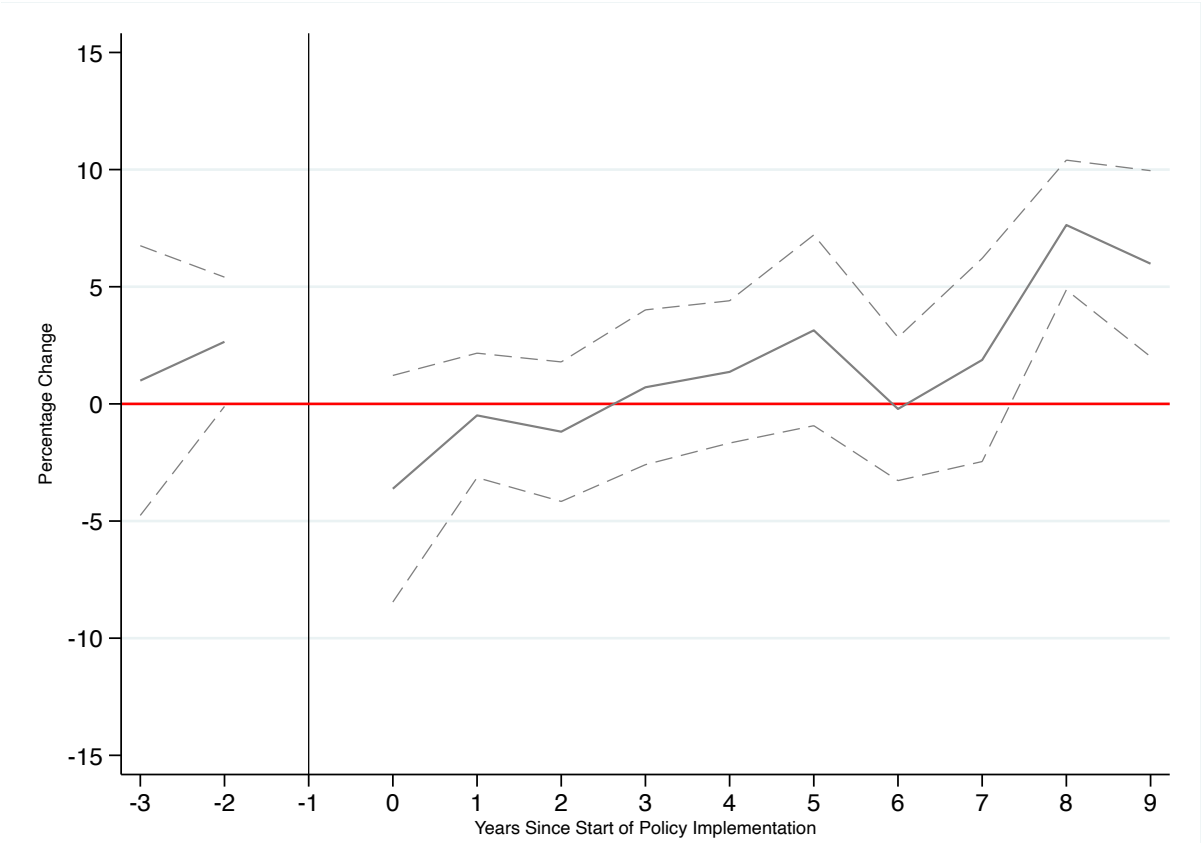
Notes. This figure presents average achievement gaps between students with and without disabilities on elementary and middle school reading and math standardized assessments between 2003 and 2014. The vertical line at 2009 marks the end of the implementation period. Elementary gap trends reflect averages across grades 3 through 5, and middle school gap trends reflect averages across grades 6 through 8. Achievement gaps represent average, group-level differences wherein each group's originally reported percent-proficient metric has been transformed into the group's average latent propensity for proficiency interpretable in a standard deviation-unit metric (Ho, 2009). Raw data are from the case study department of education.

Figure 14: High School Dropout Rates



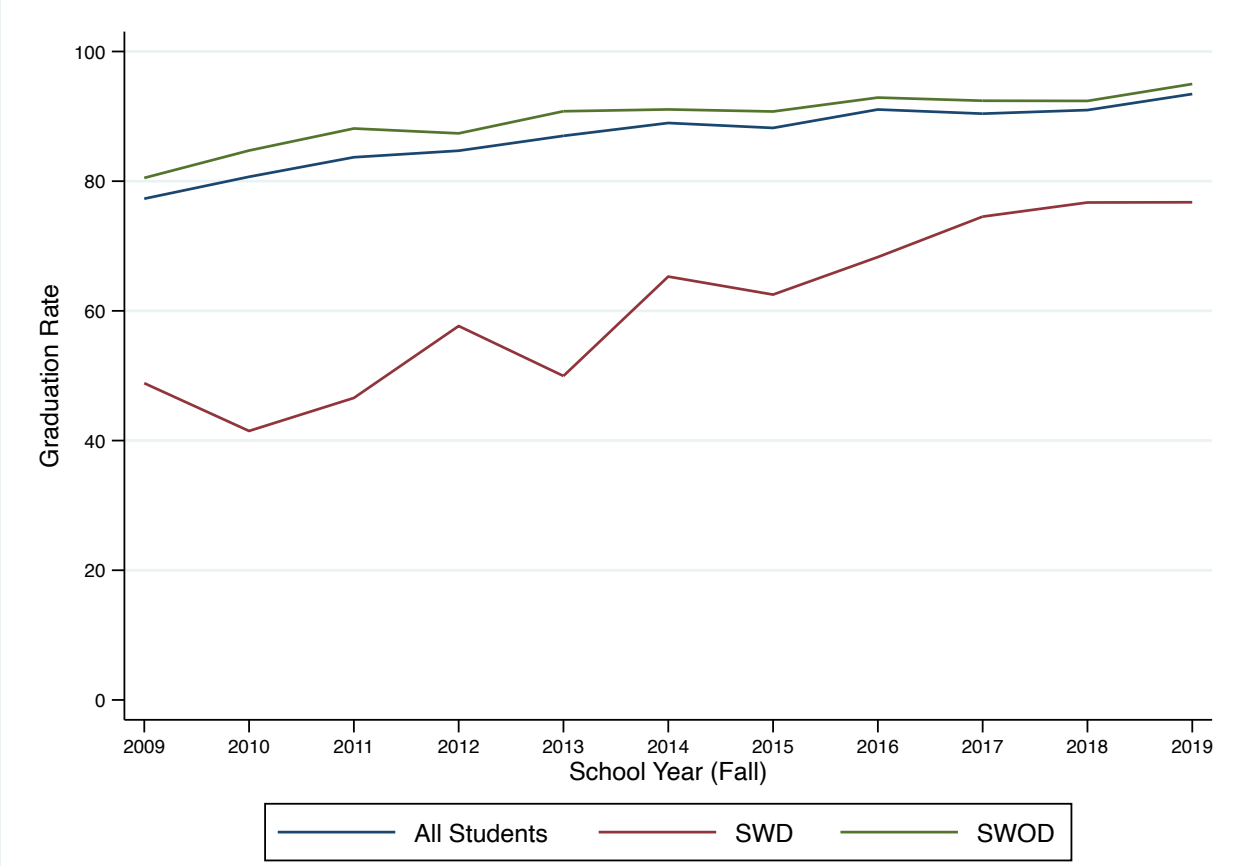
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on high school dropout rates for all students in grades 9 through 12 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 15: High School Graduation Rates



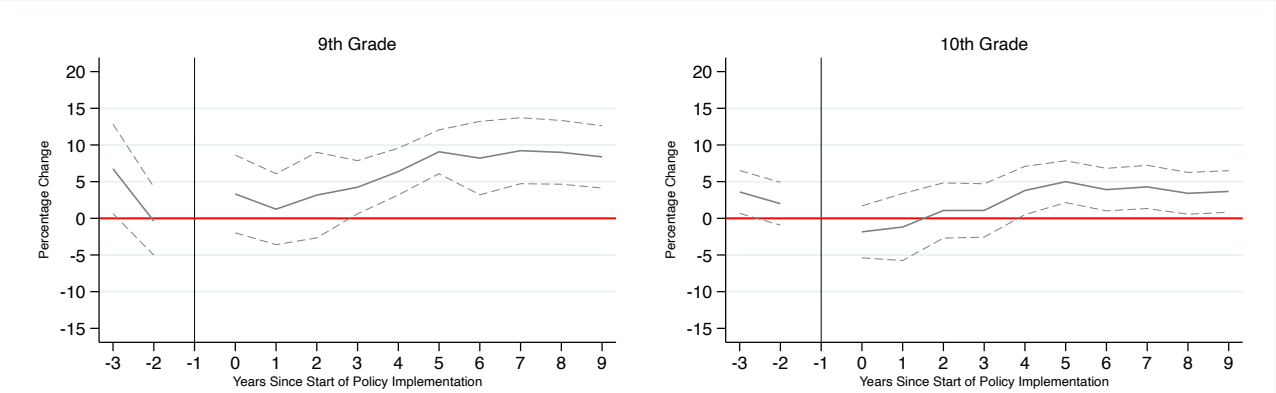
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on high school graduation rates for all students in grades 9 through 12 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 16: Post-Implementation Subgroup Graduation Rates



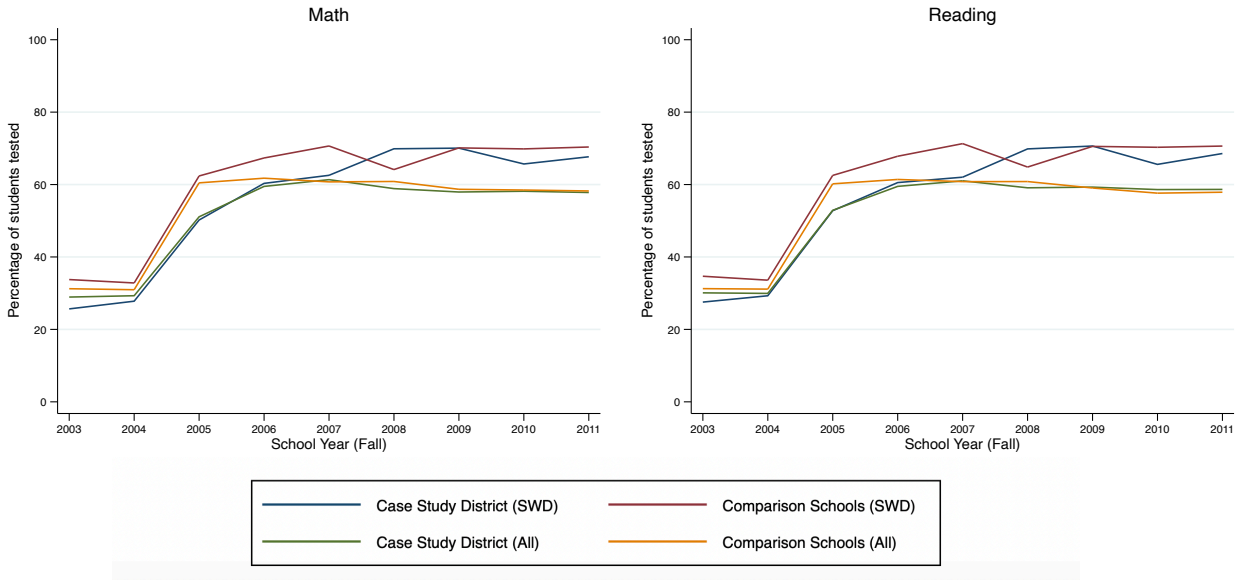
Notes. This figure presents raw, subgroup-level graduation rates for all students, students with disabilities, and students without disabilities in the case study district from 2009-2019—the post-implementation period. Data are from the case study department of education.

Figure 17: High School Promotion Rates



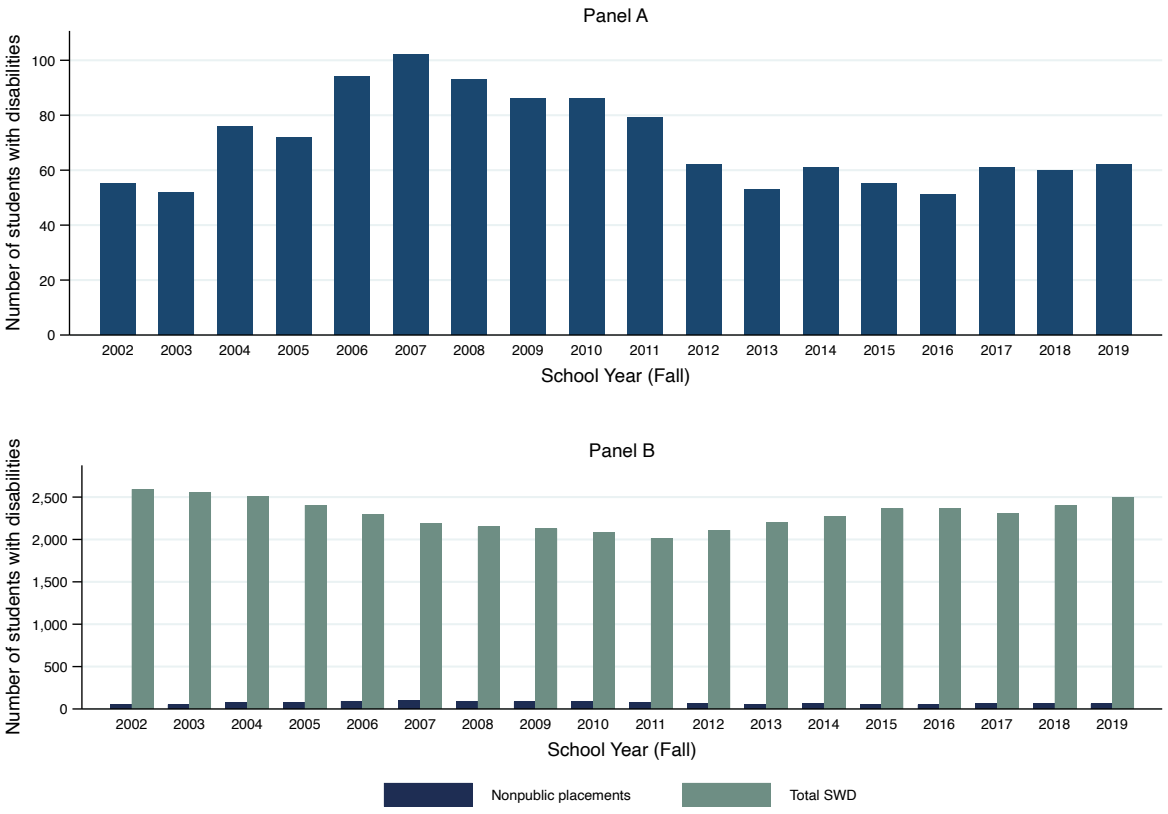
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on high school promotion rates for all students in grades 9 through 12 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group.

Figure 18: Students Tested over Time



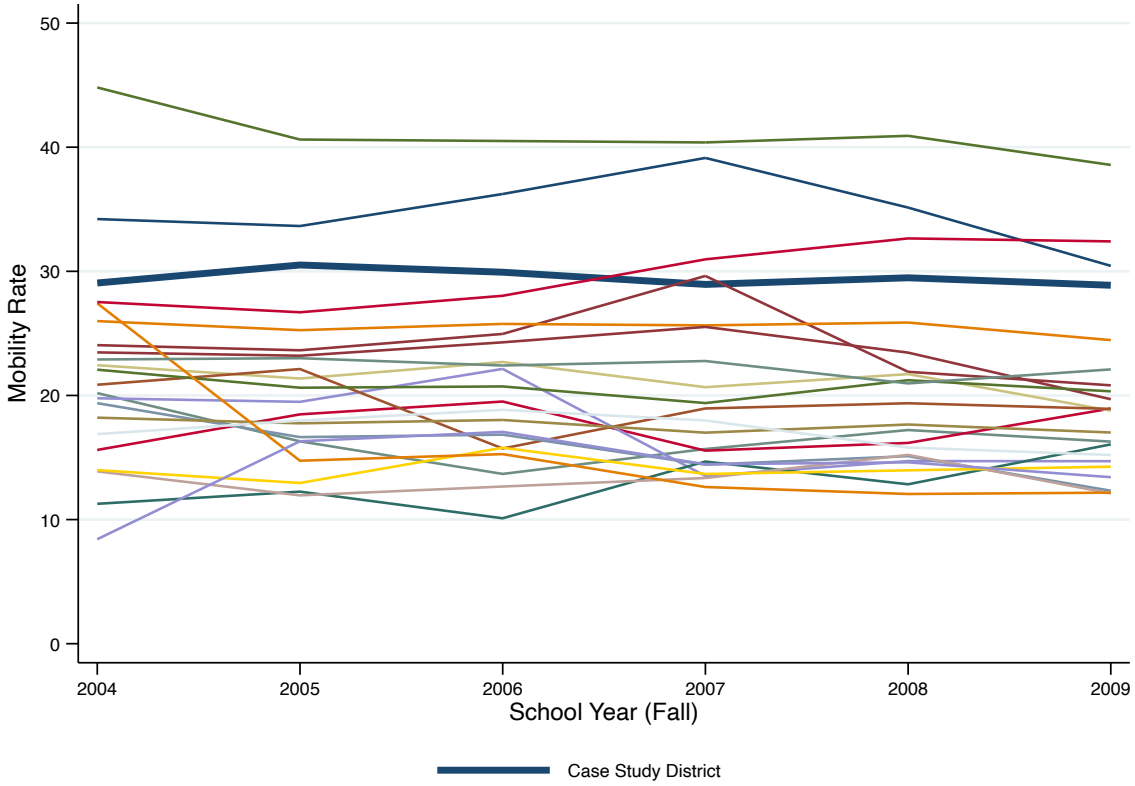
Notes. This figure illustrates the annual percentage of students participating in state standardized testing in both reading and math. Participation rates are shown for students with disabilities and all students, for both the case study district and the comparison schools (all other untreated schools within the state).

Figure 19: Nonpublic Placements



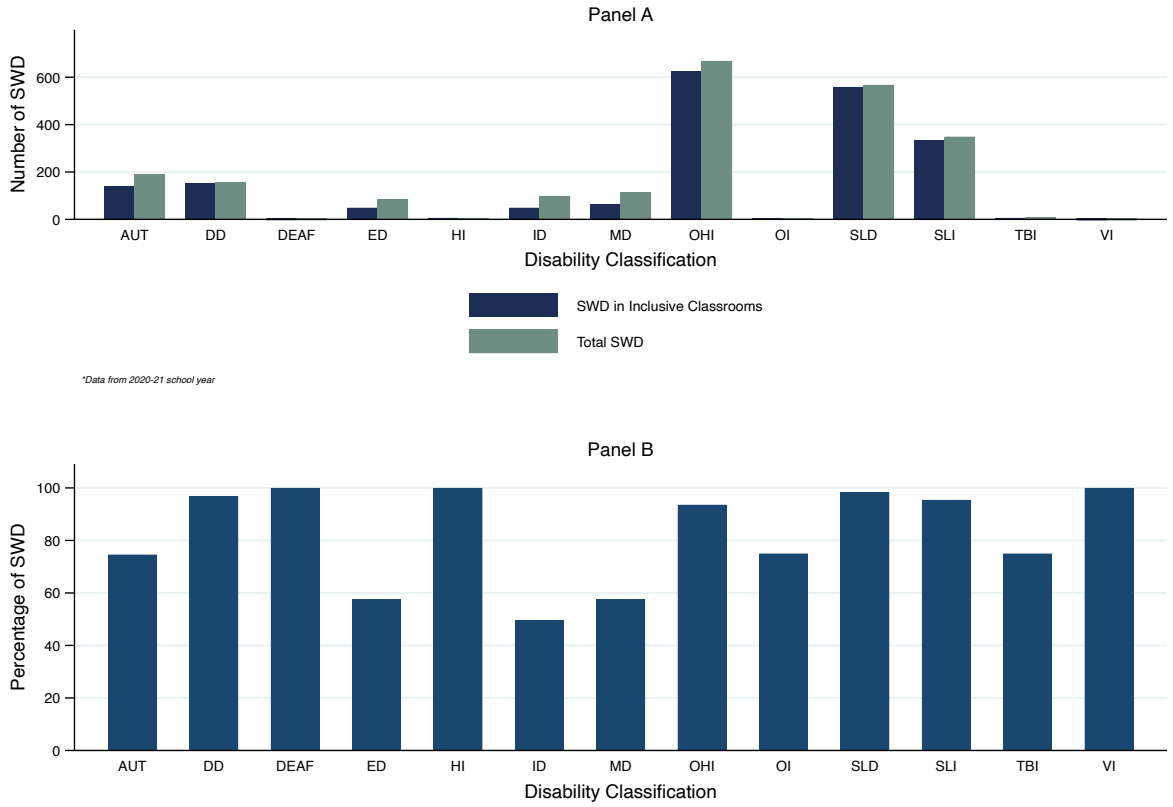
Notes. Panel A in this figure illustrates the annual number of students with disabilities in the case study district placed in nonpublic educational settings between 2002 and 2019. Panel B illustrates this same number against the total number of students with disabilities in the district each year. Data are from the case study district and the case study state department of education.

Figure 20: District Mobility Rates During Policy Period



Notes. This figure shows the annual mobility rate for the case study district alongside rates for all other districts in the state between 2004 and 2009 (the policy implementation period). Mobility rates are calculated based on the number of students voluntarily exiting district schools for reasons including: transfers to homeschooling, transfers to nonpublic schools, transfers to other public schools out-of-district, or transfers to schools in foreign countries. Data are from the case study state department of education.

Figure 21: Students with Disabilities in Inclusive Classrooms by Classification



Notes. Panel A of this figure shows the number of students with disabilities in the case study district within each of 13 disability classifications placed in general education classrooms for 80% or more of the school day as their primary educational placement (i.e., in an inclusive setting). These categories include: autism (AUT), developmental delay (DD), deafness (DEAF), emotional disturbance (ED), hearing impairment (HI), intellectual disability (ID), multiple disabilities (MD), other health impairment (OHI), orthopedic impairment (OI), specific learning disability (SLD), speech or language impairment (SLI), traumatic brain injury (TBI), and visual impairment (VI). Panel A also shows the total number of students with disabilities in the case study district in each of these classifications. Panel B reflects the same information in Panel A but in percentage form—illustrating the proportion of students with disabilities in each classification category in the case study district placed in inclusive classrooms. Student-level placement data are from the case study district for the 2020-21 school year.

Appendix A: Event Study Results Based on Alternative Comparison Groups

Comparison Group B: All Untreated Rural Schools

Table A.1: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.0188	-0.642	0.195	0.141	-0.583	0.284
	(0.242)	(0.424)	(0.320)	(0.258)	(0.438)	(0.329)
Pre-treatment mean	92.8	91.1	93	92.8	91.1	93
Number of schools	127	122	124	127	122	124
Math Test Score Means	0.0533	0.0234	0.0966*	0.0407	-0.000982	0.0744
	(0.0448)	(0.0898)	(0.0562)	(0.0500)	(0.0894)	(0.0634)
Pre-treatment mean	0.02	-0.023	0.029	0.02	-0.023	0.029
Number of schools	91	91	91	91	91	91
Reading Test Score Means	-0.00745	0.0556	0.0305	-0.000517	0.00889	0.0410
	(0.0368)	(0.0719)	(0.0419)	(0.0487)	(0.0793)	(0.0535)
Pre-treatment mean	-0.008	-0.069	-0.014	-0.008	-0.069	-0.014
Number of schools	91	91	91	91	91	91

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across grades 3 through 12 for attendance and 3 through 8 for math and reading test scores. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated rural schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table A.2: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.121	-0.218	0.0288	-0.114	-0.667***	-0.100
	(0.0744)	(0.190)	(0.0711)	(0.102)	(0.206)	(0.0686)
Pre-treatment mean	94.5	94.5	95	94.5	94.5	95
Number of schools	82	82	82	82	82	82
Math Test Score Means	0.0671	0.0190	0.120	0.0417	-0.0133	0.0840
	(0.0611)	(0.129)	(0.0778)	(0.0642)	(0.123)	(0.0861)
Pre-treatment mean	0.01	-0.041	-0.023	0.01	-0.041	-0.023
Number of schools	73	73	73	73	73	73
Reading Test Score Means	-0.0195	0.0501	0.0288	-0.0171	-0.00526	0.0359
	(0.0526)	(0.112)	(0.0606)	(0.0665)	(0.117)	(0.0758)
Pre-treatment mean	-0.030	-0.119	-0.035	-0.030	-0.119	-0.035
Number of schools	73	73	73	73	73	73

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated rural schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A.3: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.0382	-0.294	-0.0113	0.517**	0.569	0.466
	(0.135)	(0.422)	(0.149)	(0.219)	(0.610)	(0.278)
Pre-treatment mean	93.6	90.4	94	93.6	90.4	94
Number of schools	34	31	32	34	31	32
Math Test Score Means	0.0320	0.0426	0.0546	0.0589	0.0351	0.0735
	(0.0616)	(0.107)	(0.0787)	(0.0844)	(0.131)	(0.0921)
Pre-treatment mean	0.049	0.026	0.044	0.049	0.026	0.044
Number of schools	22	18	22	22	18	22
Reading Test Score Means	0.0166	0.0738	0.0325	0.0410	0.0593	0.0580
	(0.0432)	(0.0531)	(0.0500)	(0.0651)	(0.0789)	(0.0632)
Pre-treatment mean	0.053	0.072	0.047	0.053	0.072	0.047
Number of schools	22	22	22	22	22	22

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated rural schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A.4: Event Study Results (High)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.208	-1.123*	-0.0157	-0.0701	-1.321*	0.0187
	(0.414)	(0.565)	(0.397)	(0.433)	(0.679)	(0.414)
Pre-treatment mean	90.6	88.9	90.9	90.6	88.9	90.9
Number of schools	25	22	22	25	22	22
Dropout Rate	0.553	2.347***	-0.0899	-0.123	1.296*	-0.477
	(0.538)	(0.792)	(0.404)	(0.481)	(0.713)	(0.329)
Pre-treatment mean	4.2	9.9	3.3	4.2	9.9	3.3
Number of schools	27	25	25	27	25	25
Graduation Rate	-3.057*	-	-	-2.387	-	-
	(1.698)			(1.571)		
Pre-treatment mean	81.3			81.3		
Number of schools	20			20		
Promotion Rate *(All Grades)*	-3.979**	-	-	-2.988	-	-
	(1.903)			(1.869)		
Pre-treatment mean	91.3			91.3		
Number of schools	23			23		
9th Grade	0.332	-	-	2.237	-	-
	(1.294)			(2.070)		
Pre-treatment mean	87.7			87.7		
Number of schools	23			23		
10th Grade	-2.884**	-	-	-0.919	-	-
	(1.086)			(1.285)		
Pre-treatment mean	89.3			89.3		
Number of schools	20			20		
11th Grade	-2.130**	-	-	-2.165**	-	-
	(0.971)			(0.866)		
Pre-treatment mean	93.4			93.4		
Number of schools	19			19		
12th Grade	0.264	-	-	-0.465	-	-
	(0.447)			(0.614)		
Pre-treatment mean	94.6			94.6		
Number of schools	21			21		

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 9 through 12. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the

start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated rural schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Comparison Group C: Synthetic Comparison Group

Table A.5: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.250	-0.216	0.308	0.389**	-0.127	0.463**
	(0.186)	(0.402)	(0.222)	(0.194)	(0.349)	(0.234)
Pre-treatment mean	92.8	91.1	93	92.8	91.1	93
Number of schools	1080	1078	1049	1080	1078	1049
Math Test Score Means	0.0396	-0.0141	0.0229	0.0336	-0.0135	0.0272
	(0.0431)	(0.0332)	(0.0858)	(0.0447)	(0.0418)	(0.0850)
Pre-treatment mean	0.02	-0.023	0.029	0.02	-0.023	0.029
Number of schools	848	848	838	848	848	838
Reading Test Score Means	0.0519	0.0848	0.0228	0.0281	0.0654	0.0250
	(0.0672)	(0.0537)	(0.0378)	(0.0749)	(0.0586)	(0.0463)
Pre-treatment mean	-0.008	-0.069	-0.014	-0.008	-0.069	-0.014
Number of schools	847	841	841	847	841	841

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across grades 3 through 12 for attendance and 3 through 8 for math and reading test scores. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises a synthetic, data-driven composite group drawn from the full pool of untreated schools within the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A.6: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.00628	-0.170	0.124	-0.0876	-0.652***	-0.0908
	(0.0562)	(0.170)	(0.0884)	(0.0736)	(0.235)	(0.0582)
Pre-treatment mean	94.5	94.5	95	94.5	94.5	95
Number of schools	705	1,117	1,117	705	1,117	1,117
Math Test Score Means	0.0607	-0.0166	0.00893	0.0515	-0.0146	0.0112
	(0.0586)	(0.0479)	(0.126)	(0.0586)	(0.0593)	(0.118)
Pre-treatment mean	0.01	-0.041	-0.023	0.01	-0.041	-0.023
Number of schools	644	644	643	644	644	643
Reading Test Score Means	0.0629	0.117	0.0302	0.0462	0.0934	0.0345
	(0.104)	(0.0746)	(0.0557)	(0.112)	(0.0813)	(0.0683)
Pre-treatment mean	-0.030	-0.119	-0.035	-0.030	-0.119	-0.035
Number of schools	643	642	642	643	642	642

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises a synthetic, data-driven composite group drawn from the full pool of untreated schools within the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A.7: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.0751	-0.164	0.0842	0.541***	0.394	0.559**
	(0.121)	(0.313)	(0.134)	(0.164)	(0.460)	(0.235)
Pre-treatment mean	93.6	90.4	94	93.6	90.4	94
Number of schools	290	289	263	290	289	263
Math Test Score Means	0.00423	-0.0144	0.0435	0.00417	-0.0190	0.0569
	(0.0597)	(0.0423)	(0.0953)	(0.0697)	(0.0509)	(0.123)
Pre-treatment mean	0.049	0.026	0.044	0.049	0.026	0.044
Number of schools	348	348	289	348	348	289
Reading Test Score Means	0.0362	0.0315	0.00757	0.00188	0.0195	0.00249
	(0.0498)	(0.0745)	(0.0486)	(0.0699)	(0.0769)	(0.0476)
Pre-treatment mean	0.053	0.072	0.047	0.053	0.072	0.047
Number of schools	343	343	343	343	343	343

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises a synthetic, data-driven composite group drawn from the full pool of untreated schools within the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table A.8: Event Study Results (High)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.142	-1.234***	-0.0940	0.497	-0.346	0.535*
	(0.273)	(0.357)	(0.265)	(0.316)	(0.376)	(0.316)
Pre-treatment mean	90.6	88.9	90.9	90.6	88.9	90.9
Number of schools	208	205	179	208	205	179
Dropout Rate	0.124	3.061***	0.307	-0.680*	1.062	-0.222
	(0.364)	(0.839)	(0.543)	(0.385)	(0.980)	(0.484)
Pre-treatment mean	4.2	9.9	3.3	4.2	9.9	3.3
Number of schools	212	211	201	212	211	201
Graduation Rate	-1.156	-	-	2.942***	-	-
	(1.330)			(1.127)		
Pre-treatment mean	81.3			81.3		
Number of schools	191			191		
Promotion Rate *(All Grades)*	0.899	-	-	3.408***	-	-
	(1.143)			(1.241)		
Pre-treatment mean	91.3			91.3		
Number of schools	214			214		
9th Grade	1.852*	-	-	6.743***	-	-
	(1.038)			(1.347)		
Pre-treatment mean	87.7			87.7		
Number of schools	196			196		
10th Grade	-1.407	-	-	1.784*	-	-
	(1.098)			(0.972)		
Pre-treatment mean	89.3			89.3		
Number of schools	190			190		
11th Grade	0.121	-	-	0.888	-	-
	(0.813)			(0.672)		
Pre-treatment mean	93.4			93.4		
Number of schools	200			200		
12th Grade	0.493	-	-	0.821	-	-
	(0.817)			(0.884)		
Pre-treatment mean	94.6			94.6		
Number of schools	200			200		

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 9 through 12. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the

start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises a synthetic, data-driven composite group drawn from the full pool of untreated schools within the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Appendix B: Event Study Results Using Interaction-Weighted Estimator

Table B.1: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.058	-0.104	0.062	0.225***	0.025	0.239***
	0.067	0.187	0.066	0.067	0.079	0.071
Pre-treatment mean	92.8	91.1	93	92.8	91.1	93
Number of schools	1,606	1,579	1,562	1,606	1,579	1,562
Math Test Score Means	0.02	-0.023	0.064	0.008	-0.027	0.038
	0.029	0.067	0.039	0.027	0.058	0.04
Pre-treatment mean	0.02	-0.023	0.029	0.02	-0.023	0.029
Number of schools	1,057	1,043	1,050	1,057	1,043	1,050
Reading Test Score Means	-0.025	0.029	0.01	-0.028	-0.0034	0.009
	0.022	0.05	0.027	0.026	0.05	0.032
Pre-treatment mean	-0.008	-0.069	-0.014	-0.008	-0.069	-0.014
Number of schools	1,057	1,055	1,050	1,057	1,055	1,050

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across grades 3 through 12 for attendance and 3 through 8 for math and reading test scores. Estimates are adjusted to account for cohort weights using an interaction-weighted estimator (Sun & Abraham, 2021). Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table B.2: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.016	-0.07	0.071	-0.067	-0.64***	-0.118**
	0.041	0.154	0.056	0.05	0.172	0.055
Pre-treatment mean	94.5	94.5	95	94.5	94.5	95
Number of schools	1,028	1,505	1,503	1,028	1,505	1,503
Math Test Score Means	0.042	-0.039	0.092	0.025	-0.042	0.061
	0.035	0.085	0.048	0.03	0.071	0.049
Pre-treatment mean	0.01	-0.041	-0.023	0.01	-0.041	-0.023
Number of schools	799	797	797	799	797	797
Reading Test Score Means	-0.023	0.009	0.018	-0.025	-0.019	0.019
	0.027	0.061	0.033	0.033	0.056	0.04
Pre-treatment mean	-0.030	-0.119	-0.035	-0.030	-0.119	-0.035
Number of schools	799	797	797	799	797	797

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. Estimates are adjusted to account for cohort weights using an interaction-weighted estimator (Sun & Abraham, 2021). Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table B.3: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.186	-0.491***	-0.176	0.175**	-0.004	0.196
	0.101	0.185	0.098	0.078	0.18	0.117
Pre-treatment mean	93.6	90.4	94	93.6	90.4	94
Number of schools	493	480	457	493	480	457
Math Test Score Means	-0.038	0.019	-0.013	-0.05	0.018	-0.036
	0.024	0.055	0.03	0.032	0.061	0.031
Pre-treatment mean	0.049	0.026	0.044	0.049	0.026	0.044
Number of schools	451	374	446	451	374	446
Reading Test Score Means	-0.036	0.079	-0.02	-0.045	0.049	-0.029
	0.025	0.044	0.028	0.033	0.057	0.032
Pre-treatment mean	0.053	0.072	0.047	0.053	0.072	0.047
Number of schools	451	447	446	451	447	446

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. Estimates are adjusted to account for cohort weights using an interaction-weighted estimator (Sun & Abraham, 2021). Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table B.4: Event Study Results (High)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.115	-0.993***	0.189	1.06***	0.096	1.11***
	0.255	0.337	0.253	0.322	0.377	0.329
Pre-treatment mean	90.6	88.9	90.9	90.6	88.9	90.9
Number of schools	353	332	316	353	332	316
Dropout Rate	0.224	2.51***	-0.383	-0.551	0.648	-0.826***
	0.389	0.46	0.332	0.427	0.519	0.301
Pre-treatment mean	4.2	9.9	3.3	4.2	9.9	3.3
Number of schools	369	347	341	369	347	341
Graduation Rate	-1.75			2.63**		
	1.3	-	-	1.27	-	-
Pre-treatment mean	81.3			81.3		
Number of schools	325			325		
Promotion Rate *(All Grades)*	0.451			2.95***		
	0.92	-	-	1.02	-	-
Pre-treatment mean	91.3			91.3		
Number of schools	379			379		
9th Grade	1.65			6.74***		
	0.959	-	-	1.3	-	-
Pre-treatment mean	87.7			87.7		
Number of schools	352			352		
10th Grade	-1.24			2.19***		
	0.973	-	-	0.838	-	-
Pre-treatment mean	89.3			89.3		
Number of schools	336			336		
11th Grade	-0.17			0.983		
	0.804	-	-	0.675	-	-
Pre-treatment mean	93.4			93.4		
Number of schools	340			340		
12th Grade	-0.249			0.095		
	0.607	-	-	0.751	-	-
Pre-treatment mean	94.6			94.6		
Number of schools	332			332		

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 9 through 12. Estimates are adjusted to account for cohort weights using an interaction-weighted estimator (Sun & Abraham, 2021). Disaggregated outcome data are from the case study state department of education. All regressions control for school

and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Appendix C: Event Study Results Excluding PARCC Overlap Years

Table C.1: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.340*	-0.102	0.355*	0.423**	-0.159	0.448**
	(0.179)	(0.426)	(0.198)	(0.206)	(0.430)	(0.222)
Pre-treatment mean	92.8	91.1	93	92.8	91.1	93
Number of schools	1606	1579	1562	1606	1579	1562
Math Test Score Means	0.0319	0.00942	0.0753	0.00640	-0.0194	0.0364
	(0.0436)	(0.0874)	(0.0539)	(0.0505)	(0.0942)	(0.0635)
Pre-treatment mean	0.02	-0.023	0.029	0.02	-0.023	0.029
Number of schools	1057	1043	1050	1057	1043	1050
Reading Test Score Means	-0.0179	0.0509	0.0152	-0.0269	0.00887	0.00721
	(0.0324)	(0.0671)	(0.0378)	(0.0429)	(0.0821)	(0.0469)
Pre-treatment mean	-0.008	-0.069	-0.014	-0.008	-0.069	-0.014
Number of schools	1,057	1,055	1,050	1,057	1,055	1,050

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across grades 3 through 12 for attendance and 3 through 8 for math and reading test scores. The time horizon excludes years beyond 2012 to account for overlap with PARCC implementation. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table C.2: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.0235	-0.0656	0.0735	-0.163*	-0.783***	-0.161**
	(0.0609)	(0.158)	(0.0593)	(0.0965)	(0.216)	(0.0652)
Pre-treatment mean	94.5	94.5	95	94.5	94.5	95
Number of schools	1,028	1,505	1,503	1,028	1,505	1,503
Math Test Score Means	0.0560	0.00333	0.110	0.0314	-0.0278	0.0724
	(0.0600)	(0.130)	(0.0746)	(0.0672)	(0.133)	(0.0863)
Pre-treatment mean	0.01	-0.041	-0.023	0.01	-0.041	-0.023
Number of schools	799	797	797	799	797	797
Reading Test Score Means	-0.0213	0.0640	0.0212	-0.0268	0.0298	0.0175
	(0.0459)	(0.102)	(0.0531)	(0.0589)	(0.116)	(0.0643)
Pre-treatment mean	-0.030	-0.119	-0.035	-0.030	-0.119	-0.035
Number of schools	799	797	797	799	797	797

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. The time horizon excludes years beyond 2012 to account for overlap with PARCC implementation. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table C.3: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	-0.0357	-0.0299	-0.0845	0.114	0.244	0.0580
	(0.124)	(0.217)	(0.136)	(0.216)	(0.365)	(0.293)
Pre-treatment mean	93.6	90.4	94	93.6	90.4	94
Number of schools	493	480	457	493	480	457
Math Test Score Means	-0.00664	0.0220	0.0192	-0.0431	0.00770	-0.0326
	(0.0590)	(0.0945)	(0.0733)	(0.0638)	(0.122)	(0.0710)
Pre-treatment mean	0.049	0.026	0.044	0.049	0.026	0.044
Number of schools	451	374	446	451	374	446
Reading Test Score Means	-0.0164	0.0352	0.00193	-0.0400	-0.0243	-0.0232
	(0.0423)	(0.0515)	(0.0513)	(0.0483)	(0.0910)	(0.0576)
Pre-treatment mean	0.053	0.072	0.047	0.053	0.072	0.047
Number of schools	451	447	446	451	447	446

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. The time horizon excludes years beyond 2012 to account for overlap with PARCC implementation. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table C.4: Event Study Results (High)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Attendance Rate	0.114	-1.162***	0.208	1.186***	0.0651	1.237***
	(0.252)	(0.296)	(0.250)	(0.302)	(0.420)	(0.303)
Pre-treatment mean	90.6	88.9	90.9	90.6	88.9	90.9
Number of schools	353	332	316	353	332	316
Dropout Rate	0.195	2.555***	-0.400	-0.561	0.928	-0.817**
	(0.388)	(0.453)	(0.335)	(0.416)	(0.577)	(0.325)
Pre-treatment mean	4.2	9.9	3.3	4.2	9.9	3.3
Number of schools	369	347	341	369	347	341
Graduation Rate	-1.793	-	-	3.389**	-	-
	(1.293)			(1.493)		
Pre-treatment mean	81.3			81.3		
Number of schools	325			325		
Promotion Rate *(All Grades)*	0.463	-	-	3.353***	-	-
	(0.912)			(0.945)		
Pre-treatment mean	91.3			91.3		
Number of schools	379			379		
9th Grade	1.705*	-	-	6.764***	-	-
	(0.934)			(1.358)		
Pre-treatment mean	87.7			87.7		
Number of schools	352			352		
10th Grade	-1.182	-	-	2.574***	-	-
	(0.948)			(0.879)		
Pre-treatment mean	89.3			89.3		
Number of schools	336			336		
11th Grade	-0.0421	-	-	1.210*	-	-
	(0.798)			(0.730)		
Pre-treatment mean	93.4			93.4		
Number of schools	340			340		
12th Grade	-0.278	-	-	-0.0107	-	-
	(0.578)			(0.682)		
Pre-treatment mean	94.6			94.6		
Number of schools	332			332		

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 9 through 12. The time horizon excludes years beyond 2012 to account for overlap with PARCC implementation. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the

policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Appendix D: Event Study Results for Transformed Proficiency Rates with Cell Count Limitations

Table D.1: Event Study Results (All Grades)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Math Test Score Means	0.0440	0.0908	0.0900	0.0336	0.0234	0.0698
	(0.0429)	(0.117)	(0.0594)	(0.0445)	(0.162)	(0.0632)
Pre-treatment mean	0.02	0.278	0.025	0.02	0.278	0.025
Number of schools	1053	746	1042	1053	746	1042
Reading Test Score Means	-0.0114	0.0724	0.0233	-0.0141	-0.0468	0.0240
	(0.0331)	(0.0862)	(0.0377)	(0.0416)	(0.0998)	(0.0461)
Pre-treatment mean	-0.008	0.145	-0.014	-0.008	0.145	-0.014
Number of schools	1052	731	1042	1052	731	1042

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 8. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit (HOMOP). HOMOP transformations are limited to cells with student counts greater than or equal to 50 (Reardon, 2018). Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table D.2: Event Study Results (Elementary)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Math Test Score Means	0.0671	-0.0336	0.120	0.0527	-0.173	0.0921
	(0.0584)	(0.184)	(0.0745)	(0.0584)	(0.267)	(0.0812)
Pre-treatment mean	0.01	0.378	-0.023	0.01	0.378	-0.023
Number of schools	797	524	795	797	524	795
Reading Test Score Means	-0.0128	0.0530	0.0313	-0.0148	-0.0812	0.0335
	(0.0477)	(0.125)	(0.0556)	(0.0591)	(0.131)	(0.0682)
Pre-treatment mean	-0.030	0.78	-0.035	-0.030	0.78	-0.035
Number of schools	797	517	795	797	517	795

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 3 through 5. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit (HOMOP). HOMOP transformations are limited to cells with student counts greater than or equal to 50 (Reardon, 2018). Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

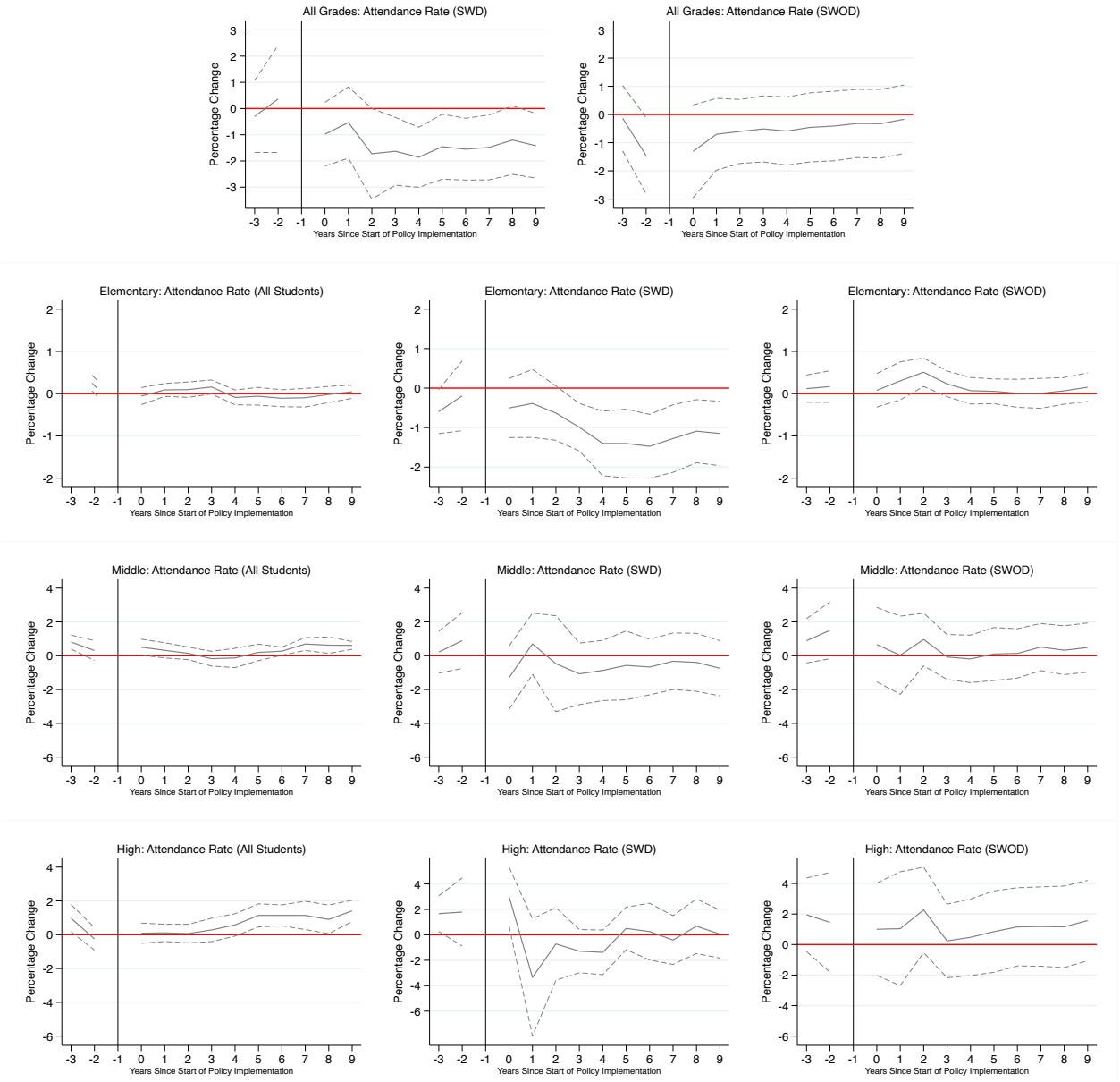
Table D.3: Event Study Results (Middle)

	Implementation Period (0-4 years)			Post-Implementation (5+ years)		
	All Students	*SWD*	*SWOD*	*All Students*	*SWD*	*SWOD*
Math Test Score Means	0.00626	0.137	0.0401	0.00162	0.146	0.0391
	(0.0589)	(0.146)	(0.0992)	(0.0686)	(0.171)	(0.100)
Pre-treatment mean	0.049	0.139	0.031	0.049	0.139	0.031
Number of schools	448	287	439	448	287	439
Reading Test Score Means	-0.0131	0.0850	0.00728	-0.0201	-0.0194	0.00197
	(0.0422)	(0.109)	(0.0485)	(0.0506)	(0.138)	(0.0475)
Pre-treatment mean	0.053	0.261	0.047	0.053	0.261	0.047
Number of schools	447	293	439	447	293	439

Notes. Coefficients are estimates of the impact of the inclusion policy's implementation on the indicated outcome for all students, students with disabilities (SWD), and students without disabilities (SWOD) across all grades 6 through 8. Disaggregated outcome data are from the case study state department of education. All regressions control for school and year-by-school-level fixed effects. Results are split into three periods: the pre-policy period (all years prior to the start of implementation), the policy implementation period (0-4 years), and the post-implementation period (5 to 9 years after the policy implementation concluded). The pre-policy period is omitted as a reference group. Leads greater than three years prior to implementation and lags more than nine years after the start of implementation were binned, respectively. Math and reading test score means are transformed from reported subgroup proficiency rates into recovered test score means using homoskedastic ordered probit (HOMOP). HOMOP transformations are limited to cells with student counts greater than or equal to 50 (Reardon, 2018). Standard errors are clustered at the school level. The sample includes all schools in the case study district, while the comparison group comprises all untreated schools in the state. Robust standard errors are in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

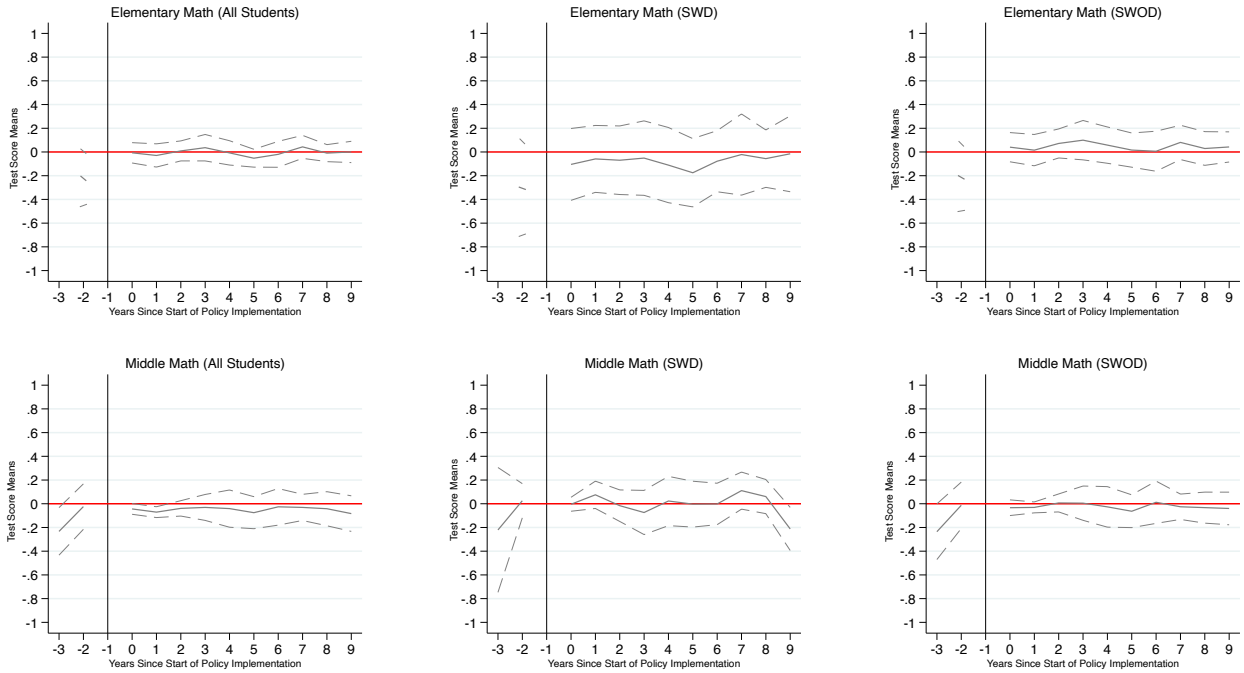
Appendix E: Additional Event Study Graphs

Figure E.1: Attendance Rates



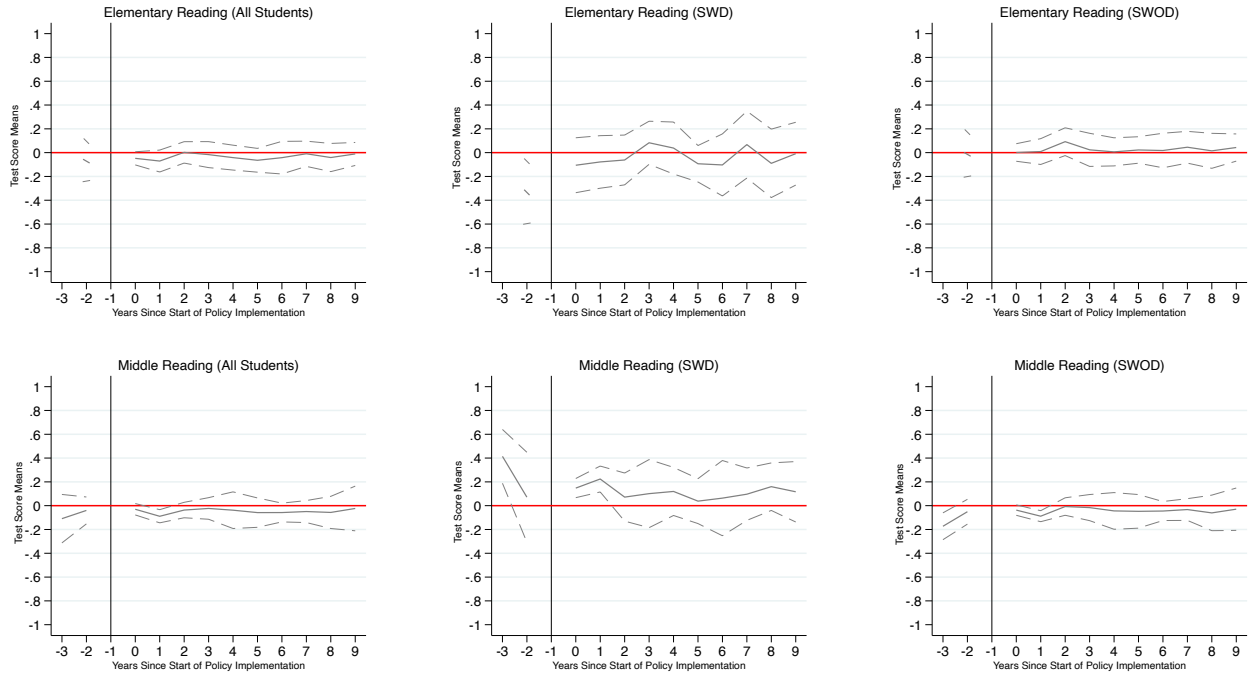
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on attendance rates for all students, students with disabilities (SWD), and students without disabilities (SWOD) in grades 3 through 12 from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group. Impacts on attendance for all students across all grades 3 through 12 is presented in the main text.

Figure E.2: Math Test Score Means



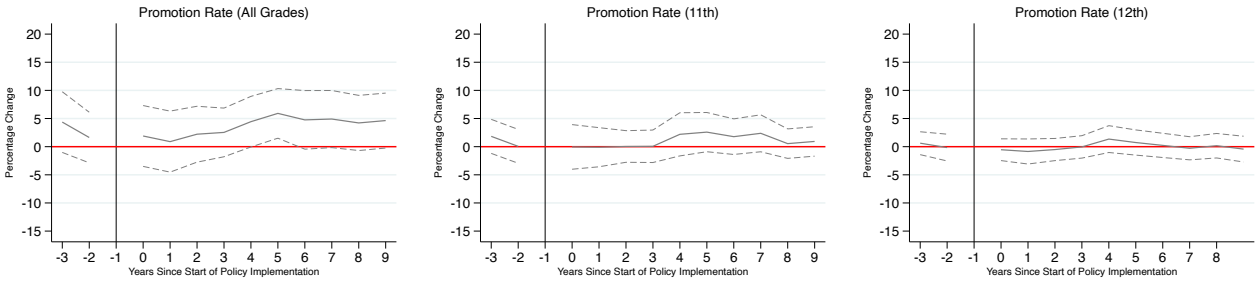
Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on math test score means for all students, students with disabilities (SWD), and students without disabilities (SWOD) in elementary school (grades 3 through 5) and middle school (grades 6 through 8) from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group. Impacts on math test score means for all students across grades 3 through 8 is presented in the main text.

Figure E.3: Reading Test Score Means



Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on reading test score means for all students, students with disabilities (SWD), and students without disabilities (SWOD) in elementary school (grades 3 through 5) and middle school (grades 6 through 8) from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group. Impacts on reading test score means for all students across grades 3 through 8 is presented in the main text.

Figure E.4: High School Promotion Rates



Notes. This figure presents results from the event study, illustrating the impact of the inclusion policy on high school promotion rates for all students in grades 9 through 12, students in 11th grade, and students in 12th grade from three years prior to the start of implementation to nine years after implementation began. The year before the start of implementation is excluded as the reference group. Impacts on promotion rates for students in 9th grade and students in 10th grade are presented in the main text.