The Effects of Early Literacy Policies on Student Achievement

June 2023

Education Policy Innovation Collaborative

COLLEGE OF EDUCATION | MICHIGAN STATE UNIVERSITY

236 ERICKSON HALL, 620 FARM LANE, EAST LANSING, MI 48824 | www.EPICedpolicy.org

ACKNOWLEDGEMENTS

We thank Scott Imberman, Melinda Morrill, Katharine Strunk, and participants at the AEFP 2022 Annual Conference, Abt Associates seminar, and workshops at Michigan State University for their helpful feedback, discussion, and comments.

DISCLAIMER

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200009 to Michigan State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

JUNE 2023

The Effects of Early Literacy Policies on Student Achievement

AUTHORS

John Westall, *EPIC, MSU*

Amy Cummings, *EPIC, MSU*

ABSTRACT

Given the importance of early literacy to long-term student success, by 2021, 41 states and the District of Columbia adopted early literacy policies to improve student literacy by the end of third grade. We use an event-study approach to examine the impact of these policies on high- and low-stakes test scores. Our results suggest that adopting an early literacy policy improves elementary students' reading achievement on high-stakes assessments, particularly in third grade and in states with comprehensive early literacy policies and third-grade retention requirements. We also find suggestive evidence that early literacy policies reduce socioeconomic and racial high-stakes achievement gaps in reading and have positive spillover effects on math achievement. However, we find little evidence of significant gains in low-stakes test scores except in states with comprehensive policies. Our findings highlight the importance of content and incentives for early literacy policies.

The Effects of Early Literacy Policies on Student Achievement

INTRODUCTION

As of 2020, only one-third of fourth and eighth graders could read proficiently, according to the National Assessment of Education Progress (NAEP) (National Center for Education Statistics, 2020). These rates declined even further over the course of the COVID-19 pandemic, with nine-year-olds' average reading performance dropping five points—the largest decline in three decades (NAEP, 2022). Children who struggle to read not only have lower academic achievement but also face adverse social and economic outcomes throughout their lives, including being more likely to drop out of school, experience mental health issues, and be incarcerated or unemployed (Cunningham & Stanovich, 1997; Daniel et al., 2006; Fiester, 2013; Fiester & Smith, 2010; Hernandez, 2011; Sparks et al., 2014).

As such, over the past two decades, policymakers have paid increasing attention to early literacy, with a specific focus on reading proficiency by the end of third grade, which is seen as a critical benchmark for further learning and later outcomes (Cunningham & Stanovich, 1997; Fiester & Smith, 2010; Hernandez, 2011; Sparks et al., 2014). By 2021, 41 states and the District of Columbia had such an early literacy policy (ExcelinEd, 2021). These policies vary in content and intensity but share many core components, including diagnostic early literacy assessments beginning in kindergarten, evidence-based literacy instruction and interventions, parental involvement, and professional development for educators, all to improve students' literacy in grades K-3.

Despite the ubiquity of early literacy policies, there is limited research on their effects on student achievement and literacy learning. Notable exceptions include longitudinal studies conducted in Florida and Michigan, which find positive short-term effects (Greene & Winters, 2004, 2006, 2007, 2009; Schwerdt et al., 2017; Strunk et al., 2021). However, to the best of our knowledge, there are no national assessments of early literacy policies across states. Therefore, we explore the following questions: (1) How do early literacy policies affect reading and math achievement on low- and high-stakes

assessments? (2) Do these effects depend on policy composition? (3) Do early literacy policies impact economic or racial test-score gaps?

This study contributes to existing research in several ways. First, prior studies focus only on the impact of early literacy policies on high-stakes assessments (Greene & Winters, 2004, 2006, 2007, 2009; Schwerdt et al., 2017; Strunk et al., 2021). While increases in high-stakes test scores could indicate improvements in students' literacy skills, they might also reflect "teaching to the test" or other policy-induced behavioral changes (Popham, 2001; Jacob, 2005). To understand the impact of early literacy policies on student learning more thoroughly, we use both high- and low-stakes assessments as outcomes. Second, prior research evaluates early literacy policies in single states, examining net impacts (Strunk et al., 2021) or the effects of specific elements such as third-grade retention (Greene & Winters, 2004, 2006, 2007, 2009; Schwerdt et al., 2017). Our national setting and detailed policy data allow us to estimate the net impact of early literacy policies nationwide, providing external validity to our findings. Moreover, variation in the content of early literacy policies across states allows us to examine whether specific policy components improve student achievement.

Our analysis uses three publicly available datasets. First, we use a state-level early literacy policy database published by ExcelinEd in 2021, which contains information on all states' early literacy policies, including their core components and dates of passage. Second, to measure high-stakes assessment outcomes, we rely on the Stanford Education Data Archive (SEDA), which provides yearly data on aggregated performance on high-stakes reading and math assessments across states from 2009 to 2018 (Reardon et al., 2021). Third, we measure low-stakes assessment outcomes using NAEP, which provides average performance on low-stakes reading and math assessments in fourth and eighth grades across states every two years from 1992 to 2019 (The Nation's Report Card, n.d.).

Due to states' staggered adoption of early literacy policies, we draw from the growing literature on robust difference-in-differences estimators (e.g., Callaway & Sant'Anna, 2021; Sun & Abraham, 2021) and use event-study models that leverage variation in the adoption and content of these policies across states and over time. We find that having an early literacy policy improves students' performance on high-stakes reading assessments in elementary school, with the largest effects appearing in third grade and diminishing thereafter. The effects appear to be driven by the inclusion of a third-grade retention mandate and comprehensive early literacy policies that include additional interventions and supports experience the strongest effects. However, we find limited evidence of improvements in low-stakes reading scores, except in states with comprehensive policies. We also provide evidence of positive spillovers on math achievement and suggestive evidence that early literacy policies reduce socioeconomic and racial high-stakes reading test-score gaps. Our results are robust to different estimators and pass falsification tests.

We also explore whether contemporaneous education policy initiatives drive the effects of early literacy policies. Most states began adopting early literacy policies in the 2010s, around the same time the U.S. Department of Education's Race to the Top (RTTT) program began providing billions of dollars in education grants to states with winning applications (The White House, 2017). RTTT incentivized states to adopt standards such as the Common Core by prioritizing them in the grant competition. The Common Core includes standards in both reading and math (Common Core State Standards Initiative, n.d.), so we may expect receiving RTTT funds to affect achievement in these subject areas. Thus, we separately analyze the states that did and did not receive RTTT funds. Our reading results appear to be attributable to early literacy policies, as states that did not receive RTTT funds still experienced significant positive effects on high-stakes scores. This is not the case for math scores. States that received RTTT funding experienced significant positive effects on high-stakes math scores, whereas those that did not receive funding did not experience significant gains. We, therefore, interpret our math results with caution.

The remainder of this paper is organized as follows. Section 2 provides an overview of early literacy policies and the literature surrounding them. We describe our data and methods in Section 3. In Section 4, we describe the results. Section 5 presents additional robustness checks. We discuss the policy implications of our findings and conclude in Section 6.

BACKGROUND AND RELEVANT LITERATURE

While most states have early literacy policies, these policies differ from one state to another. ExcelinEd (2021) recently reviewed all states' policies to determine whether they included 16 possible components under four categories: (1) supports for teachers and policy implementation, (2) assessment and parent notification, (3) instruction and intervention, and (4) retention and intensive intervention. Table 1 details each of the 16 policy components and shows the number of states whose policies incorporate them.

As ExcelinEd's (2021) dataset does not identify the years in which states passed each individual policy component, we focus on three key distinctions between states, illustrated in Figure 1. The first is between states with any early literacy policy (i.e., at least one of the 16 components listed in Table 1) and those without. Forty-one states and D.C. (82%) have an early literacy policy. The second is between states with third-grade retention mandates and those without. Twenty-two states (43%) require retention. We focus on retention because previous studies (e.g., Schwerdt et al., 2017) have relied on it to evaluate the effects of early literacy policies on student achievement, and because retention is the most controversial of the 16 policy components (Burns, 2016; Fiester, 2013; Starr, 2019). The third is between states with comprehensive early literacy policies and those without. We follow ExcelinEd (2021) in defining a comprehensive early literacy policy as one with all 16 policy components listed in Table

1. Twelve states (24%) have such a comprehensive policy. As comprehensive policies include all policy components, they are a subset of states with retention components.

As mentioned, research on early literacy policies is limited to a few single-state assessments. Greene and Winters (2004, 2006, 2007, 2009) conducted multiple quasi-experimental analyses of Florida's early literacy policy, primarily relying on discontinuities created by its retention component. They find positive short-term effects on students' reading achievement on the state assessment (and positive spillover effects on third-grade math achievement), but these effects dissipate as students progress through school (Greene & Winters, 2004, 2006, 2007, 2009; Schwerdt et al., 2017). Moreover, studies using regression discontinuity designs provide only local treatment effect estimates that are valid near the discontinuity (i.e., the cutoff score that identifies students for retention). Florida's fourth-grade reading scores on NAEP also improved under its early literacy policy, but it is unclear whether retention or other policy components drove these improvements (Duke et al., 2014).

There is also an ongoing effort to evaluate Michigan's early literacy policy (Strunk et al., 2021, 2022). Early results from this evaluation indicate that third-grade student reading achievement on Michigan's state test improved and that educators attribute gains to the literacy supports outlined in the policy (Strunk et al., 2021). Notably, researchers rely on high-stakes assessments to evaluate Michigan's and Florida's policies.

We contribute to the existing literature by providing the first known national assessment of early literacy policies and examining their effect on student achievement on high- and low-stakes tests. Evaluating high- and low-stakes outcomes is important because while increases in high-stakes test scores could indicate improvements in literacy skills, they could also reflect other phenomena like "teaching to the test" (Popham, 2001) or changes in tested content. Because states' high-stakes summative assessments often relate to standards and curricula, the material on which students are evaluated may dictate what teachers teach in their classrooms. Furthermore, states with third-grade retention requirements identify retention-eligible students using their state's high-stakes assessment, meaning they have an extra incentive to focus on tested skills. Thus, if early literacy policies also positively affect low-stakes test scores, this could provide further evidence of improved literacy skills. We also follow cohorts of students differentially exposed to early literacy policies as they progress through school. This will help determine whether the effects of early literacy policies are lasting or transient and whether more years of exposure to the policy enhances its effect.

In addition to examining the overall effects of early literacy policies, we explore the differential effects of having two early literacy policy compositions that we can identify with our data: (1) those that include third-grade retention and (2) those that are comprehensive. There is mixed evidence on the effectiveness of grade retention for short- and long-term cognitive and non-cognitive outcomes (e.g., Eren et al., 2022;

Greene & Winters, 2004, 2006, 2007; Holmes & Matthews, 1984; Hong & Yu, 2007; Jacob, 2005; Jacob & Lefgren, 2004; Jimerson, 2001; Lorence, 2014; McCombs et al., 2009; Nagaoka & Roderick, 2004; Roderick & Nagaoka, Jenny, 2005; Schwerdt et al., 2017; Weiss et al., 2018; Winters & Greene, 2012; Wu et al., 2010). Thus, including third-grade retention can either enhance or attenuate the effects of early literacy policies. Meanwhile, the other policy components included in ExcelinEd's (2021) dataset (e.g., professional development for teachers and interventions for students who need additional literacy support) have led to positive effects on student achievement (see Strunk et al., 2021 for a thorough review), so we might expect comprehensive early literacy policies that include all of these components to have stronger effects.

Early literacy policies may also improve student achievement outside of reading. On the one hand, improved literacy skills might improve students' test-taking ability because they can better comprehend test questions. On the other, increased funding for literacy interventions induced by early literacy policies may reduce funds available for other subjects. We focus on math achievement because previous research finds that Florida's early literacy policy positively affected students' test performance in this area (Greene & Winters, 2004; Winters & Greene, 2012).

Lastly, early literacy policies might affect socioeconomic and racial achievement gaps. These gaps are well established in the literature (Fryer & Levitt, 2004; Reardon et al., 2019), but whether they increase, decrease, or remain the same under early literacy policies depends both on the policy's allocation of interventions and resources and whether its effects on student outcomes are heterogeneous. Even if policymakers target interventions for historically underserved students, if the interventions are less effective at improving these students' outcomes, achievement gaps may not improve. Further, early literacy policies with third-grade retention requirements have been criticized for their potential for discrimination and disparate outcomes (Greene & Winters, 2009; Licalsi et al., 2019; Livingston & Livingston, 2002; Valencia & Villarreal, 2004), leading to potential increases in test-score gaps.

DATA AND METHODS

Data

No single dataset links early literacy policies to high- or low-stakes assessment data. We, therefore, combine data from three publicly available sources to assess whether early literacy policies improve student reading achievement: (1) ExcelinEd's (2021) early literacy policy database, (2) SEDA, and (3) NAEP.

Early Literacy Policy Data

ExcelinEd's (2021) early literacy policy database allows us to determine which states have an early literacy policy, when they first implemented it, and its content (i.e., whether it includes each of the 16 components described above). One of us

independently verified the ExcelinEd (2021) database by checking states' early literacy policy documents. Based on this, we made a change to the database. Oklahoma passed its early literacy policy in 1997 (Oklahoma State Department of Education, 2020), not in 2011, as listed in the ExcelinEd (2021) data. We say that a state is "treated" if it has at least one early literacy policy component and passed its early literacy policy before the last year of test-score data included in the analysis (2018 for the SEDA and 2019 for the NAEP). States that never implemented an early literacy policy during this time form the "never-treated" comparison group. Figure 2 shows the number of states passing early literacy policies in a given year.

Variation in policy composition and changes in early literacy policies over time introduce complexity to our analyses. States sometimes amend their early literacy policies (e.g., by adding policy components), leading to multiple passage dates listed in the ExcelinEd (2021) database. However, the dataset does not identify which components were added in later years. Therefore, we always use the earliest adoption date to define treatment status and interpret our effects as intent-to-treat effects.

SEDA Data

SEDA enables us to evaluate the effects of early literacy policies on high-stakes assessments. We use state-level average third- through eighth-grade reading and math achievement scores as well as non-economically disadvantaged–economically disadvantaged and White–Black test-score gaps from 2009 to 2018 (Reardon et al., 2021). SEDA test scores are derived from state-level high-stakes testing data from assessments states use for accountability purposes. Because states use different assessments, SEDA norms test scores to the NAEP to allow cross-state comparisons.

We also use state-by-year-level SEDA data regarding average demographic and macroeconomic conditions, including grade-level enrollment and the proportion of students by urbanicity, race and ethnicity, economic disadvantage status, English learner status, and special education status. Additionally, the data include the proportion of households with a single mother and a parent with at least a bachelor's degree, log median income, household poverty rates, and unemployment rates. We use these data to compare the characteristics of states with and without early literacy policies and to test whether differential trends in these characteristics explain the effects of early literacy policies on student achievement.

NAEP Data

NAEP allows us to assess the effect of early literacy policies on a low-stakes assessment because it is not tied to state accountability requirements or early literacy policy mandates (e.g., third-grade retention; The Nation's Report Card, n.d.). NAEP is designed to be comparable over time and across states, avoiding accountability-related high-stakes test-score inflation (Jacob, 2005). We use state-level average fourth- and eighth-grade reading and math scale scores from 2003 to 2019.[1] Although these policies aim to improve literacy in kindergarten through third grade, third-grade

students do not take the NAEP. Further, if states implement policies to improve early literacy in grades K-3, these efforts should be reflected in fourth-grade reading achievement, although the direct effects may lag by one year. We also examine eighth-grade NAEP reading scores to determine whether any effects of early literacy policies are sustained as students progress through school and fourth- and eighth-grade math scores to assess potential spillover effects.

Summary Statistics

Table 2 shows the number of states and state-by-year observations available for our analysis by grade, subject, outcome, and data source. Our analysis sample excludes any states that adopted an early literacy policy before our analysis period because these states cannot provide any identifying variation (i.e., we don't see a *change* in policy that might inform estimates of the impact of policy adoption).[2] For SEDA, the overall average reading scores include 45 states because the remaining states adopted an early literacy policy before 2009, and we exclude them from the sample. The state-by-year observations ranged from 391 to 400. For SEDA's overall math scores, the sample again includes 45 states in the third through sixth grades but only 44 states in the seventh and eighth grades. The sample sizes range from 352 to 400 state-by-year observations. While the same number of states have available test-score gap information, states reported this data less frequently. Thus, these outcomes consistently had fewer state-by-year observations than the overall average scores.[3] The NAEP data are complete for 47 states that adopted early literacy policies after 2003 in each available year, forming a balanced panel of 423 state-year observations.

Table 3 shows summary statistics for the outcome variables of interest in fourth and eighth grades —the two grade levels for which we have data from both SEDA and NAEP.[4] The first six columns show state average reading and math achievement scores from SEDA, including overall and by treatment status. SEDA test scores are measured in standard deviations from the national mean. The following six columns present state average NAEP reading and math scale scores, again overall and by treatment status. We also include White–Black and non-economically disadvantaged–economically disadvantaged student test-score gaps, available in the SEDA but not NAEP.

Overall, these summary statistics suggest that states with early literacy policies had substantially higher average reading and math scores on high-stakes state assessments than states that never passed early literacy policies from 2009 to 2018. However, treated states also had larger socioeconomic and racial test-score gaps. The low-stakes NAEP scores were nearly identical in the treated and never-treated states during this period.

Table 4 shows summary statistics for students' demographic characteristics. These data were only available in the SEDA from 2009 to 2019. Treated states have higher proportions of urban, suburban, and Black students, while never-treated states have higher proportions of town, rural, White, and Asian students. Regarding

macroeconomic conditions, treated states have higher unemployment rates and higher proportions of single-mother households, but never-treated states have higher proportions of economically disadvantaged students. Overall, this suggests that states with early literacy policies may serve a somewhat higher proportion of historically underserved students, but the two groups are not substantially different. Moreover, our empirical strategy controls for time-invariant differences across states using state fixed effects. We also test for and find no evidence of differential trends in state characteristics in Section 5.1.

Methods

We leverage the differential timing of states' initial early literacy policy passage and differences in policy content as sources of quasi-experimental variation. We employ an event-study identification strategy to estimate the effects of early literacy policies on student achievement. We use data on states' average test scores before and after the passage of an early literacy policy, as well as a comparison group of states that do not (or did not yet) have early literacy policies.

As mentioned, states passed early literacy policies at different times (see Figure 2). A growing body of literature has identified issues with using two-way fixed-effects and ordinary least squares (OLS) to estimate the average effect of the treatment on the treated (ATT) in difference-in-differences and event-study research designs in settings with staggered adoption and treatment-effect heterogeneity (e.g., Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Roth et al., 2022; Sun & Abraham, 2021). The former is inevitable, and the latter is likely in our setting. Accordingly, we estimate our models using the method proposed by Callaway and Sant'Anna (2021), wherein we estimate group-time average treatment effects for each group of states *g* treated in a given year at time *t*. The group-time average treatment effect *ATT(g,t)* is identified nonparametrically by computing

$$\mathrm{ATT}(\mathrm{g},\mathrm{t}) = (\mathrm{E}[\mathrm{Y}(\mathrm{g})_{\mathrm{t}}] - E[Y(NT)_t]) - \left(E\left[Y(g)_{g-\delta-1}\right] - E\left[Y(NT)_{g-\delta-1}\right]\right) \quad (1)$$

where $\delta \in \mathbb{N}_0$. Here $E[\mathrm{Y}(\mathrm{g})_{\mathrm{t}}]$ represents average test scores in year t for states adopting an early literacy policy in year g. $E[Y(NT)_t]$ represents average test scores in year t for never-treated states. $E\left[Y(g)_{g-\delta-1}\right]$ represents average test scores in states treated in year g in year $g-\delta-1$, which is some reference year before g. $E\left[Y(NT)_{g-\delta-1}\right]$ are the average test scores in the same reference year, $g-\delta-1$, for never-treated states. If $t \geq g$, meaning after the adoption of an early literacy policy, then $\delta = 0$, and the reference year is the year just before policy adoption, $g-1$. If $t < g$, meaning pre-early literacy policy, then $\delta = g-t$ and the reference period is the prior year, $t-1$.

We estimate $ATT(\mathrm{g},\mathrm{t})$ for each combination of treatment year g and calendar year t using Sant'Anna and Zhao's (2020) doubly robust difference-in-differences estimator based on stabilized inverse probability weighting and OLS. We cluster standard errors

at the state level. This estimator uses state fixed-effects, and we include no additional covariates.[5] These $ATT(\mathrm{g},\mathrm{t})$ estimates represent all the possible 2x2 difference-in-differences where the comparison groups are never-treated states, and the reference year is the prior year ($t-1$) pre-treatment and the year just before treatment ($g-1$) post-treatment. We then aggregate the $ATT(\mathrm{g},\mathrm{t})$ estimates to form event-study estimates weighted by the number of observations in a year relative to treatment. The aggregation is done as follows:

$$\theta_e^{event} = \sum_{g=1}^{\mathrm{T}} \mathbf{1}(g+e \leq \mathrm{T})\, ATT(g, g+e) P(G=g | G+e \leq \mathrm{T}, NT=0) \qquad (2)$$

where *e* represents the years relative to treatment, *g* represents the year of treatment (and $g=1$ indicates the first treatment year), and T is the last year of data availability. θ_e^{event} is the event-study coefficient estimate *e* years relative to treatment, and it equals a weighted average of the $ATT(g,t)$ estimates such that $t=g+e$ (the ATT estimate for *e* years relative to treatment in year *g*). The weights, $P(G=g|G+e \leq \mathrm{T}, NT=0)$, equal the probability that a state first adopted a policy in year *g*, given that the state was ever treated and that data are available. We perform this aggregation for five years before and after treatment in the SEDA, and eight years before treatment and nine years after treatment in the NAEP.[6]

Because of our reference-group selection, we can estimate event-study estimates for all years relative to treatment, unlike traditional event-study models, which must omit at least one year. Furthermore, pre-policy event-study coefficients are interpreted differently than in a standard two-way fixed-effects event study, where *g*-1 is the common reference year. Instead of positive (negative) pre-policy coefficients representing negative (positive) pre-policy trends, the Callaway and Sant'Anna (2021) pre-policy event-study coefficients have the opposite interpretation. This difference is because the reference year for the pre-policy Callaway and Sant'Anna (2021) event-study estimates is *t*-1, the year before the estimated effect, rather than *g*-1, the year before treatment onset. Thus, we interpret positive pre-treatment coefficient estimates as increases in the outcome relative to never-treated states before adopting an early literacy policy and negative pre-treatment coefficients as relative decreases.

In our SEDA results, we take the estimates from the grade-level event study models and group them to create cohort-level event studies. This means we follow a specific cohort of students, such as those in the third grade when an early literacy policy is passed, as they progress through school. For example, consider the cohort of students in third grade when an early literacy policy is passed in $t+0$. The cohort-level analysis follows this group to fourth grade in year $t+1$, fifth grade in year $t+2$, and so on until they reach eighth grade in $t+5$. The estimates for these cohort-level event studies are derived from the corresponding grade-level event studies (e.g., third grade $t+0$, fourth grade $t+1$, fifth grade $t+2$). Therefore, the point estimate from the cohort-level event study can be interpreted as the change in test scores for students in grade *g* in states with early literacy policies compared to those in never-treated states in year

$t + s$ relative to $t - 1$ (post-treatment, $t + s - 1$ pre-treatment). We present grade-level event studies of NAEP results because we only have fourth- and eighth-grade test scores, which do not provide compelling cohort-level event studies.

Identifying Assumptions

Our data must satisfy two key identifying assumptions for our estimates to have a causal interpretation. The first is the parallel trends assumption, which states that the average test scores of the states that were first treated in year *g* and the never-treated states would have followed parallel paths in the absence of treatment. We test the plausibility of this assumption in two ways. First, our event study allows us to test for differential test-score trends before early literacy policy passage. Significant pre-policy event-study coefficients could indicate a violation of this assumption. Second, we test for differential trends in state-level average demographic characteristics and macroeconomic conditions. Here, we estimate our event-study models using various demographic characteristics and macroeconomic conditions as outcomes. Importantly, we do not expect early literacy policies to affect these variables. Significant estimates in these falsification tests would suggest a violation of parallel trends. Overall, we find no substantial evidence of violations of the parallel trends assumption. The detailed results of the falsification tests are presented in Section 5.1.

The second assumption is that states do not anticipate their early literacy policies in ways that would alter their average test scores. This assumption might be violated if, for example, educators know that their state will pass an early literacy policy next year, so they change their behavior to improve students' literacy performance in anticipation. Anticipation of the policy can be detected in the pre-policy event-study estimates by looking at the estimate in the year before adoption.

RESULTS

Impacts on High-Stakes Reading Scores

We begin by examining whether early literacy policies impact high-stakes reading test scores. We then investigate whether these effects depend on policy composition, including whether the policy is comprehensive or contains a third-grade retention requirement.

Any Early Literacy Policy

Figure 3 shows event-study estimates of the ATT of early literacy policies on high-stakes reading scores for different cohorts of students relative to the year their state passed an early literacy policy. The x-axis represents years relative to treatment, where zero is the year in which the state passed its early literacy policy, negative values represent years until treatment (i.e., the number of years before the state passed its policy), and positive values represent years since treatment (i.e., the number of years since the state passed its policy).

Moving from left to right in Panel A, the first cohort of students is in third grade five years before their state passed an early literacy policy, meaning that they were in eighth grade by the time the policy was passed. The first data point for this cohort (five years pre-treatment) represents their average high-stakes reading score in third grade; the second data point represents their average score in fourth grade, and so on, through eighth grade. The next cohort in the figure is in seventh grade by the time the policy is passed, and so on, through the cohort in fourth grade when the policy is passed.

Because early literacy policies are targeted at grades K-3 with the aim of improving third-grade literacy achievement, we would not expect the policy to affect high-stakes reading scores for any of the cohorts in Panel A of Figure 3. Indeed, we find no statistically significant effects for these cohorts. Panel A also shows no evidence of differential pre-policy trends, supporting the parallel trends assumption. However, we see some positive impacts for fourth- and fifth-grade students after their state passed an early literacy policy. This is not surprising given that grades K-5 are typically served in the same school, so if an elementary school began implementing literacy efforts in conjunction with the passage of an early literacy policy, we might expect to see spillover effects in the fourth and fifth grades.

Turning to Panel B, these cohorts of students were in the third grade or a lower grade level when their state passed an early literacy policy, so we would expect their high-stakes reading scores to be affected. Indeed, we see positive, statistically significant effects for these cohorts of students on their third-, fourth-, and fifth-grade reading scores of a magnitude between 0.025 and 0.05 standard deviations in states that passed early literacy policies relative to those that did not. These effects fade out and are no longer statistically significant as cohorts age into middle school. Together, these results indicate that early literacy policies improve high-stakes reading achievement in the elementary grades.

Comprehensive Early Literacy Policies

Figure 4 explores whether there are differences in the effects of early literacy policies on high-stakes reading scores depending on whether they are comprehensive. Panel A shows estimates for cohorts of students in states with comprehensive early literacy policies, whereas Panel B shows estimates for cohorts in states where literacy policies are not comprehensive. These estimates measure both the effect of having a comprehensive early literacy policy and the "threat" of having one in the future, as some states added policy components in the years following the initial passage of their early literacy policy that later made it comprehensive (i.e., an intent-to-treat effect). In Figure 4 and subsequent figures showing reading results, we only show cohorts affected by early literacy policies (i.e., those in third grade or prior by the time an early literacy policy was passed), as results for earlier cohorts mirror those in Figure 3, Panel A.

States with comprehensive early literacy policies experience substantial and sustained increases in high-stakes reading scores following the adoption of their policies. These

estimates range from about 0.025 standard deviations for third-grade students when their state passed an early literacy policy to 0.1 standard deviations for those who started kindergarten a year after their state passed an early literacy policy. These estimates are substantially larger than the overall effect, suggesting that comprehensive policies play an important role in explaining the high-stakes reading score increases from early literacy policies. Further, unlike the overall results, the effects persist for several years after policy passage in states with comprehensive policies. We find limited evidence of increased high-stakes reading achievement in states without comprehensive policies.

Third-Grade Retention

Figure 5 explores the differential effects of early literacy policies based on whether the policy mandates retention for third graders whose state ELA test scores indicate that they are behind in reading. Similar to our analysis of comprehensive policies, the impact of having a retention component measures both the effect and threat of retention in the future (i.e., an intent-to-treat effect). Panel A shows estimates for cohorts of students in states whose early literacy policy includes a third-grade retention mandate, while Panel B shows estimates for cohorts of students in states whose policies do not include such a mandate.

Similar to the results for states with comprehensive early literacy policies, states whose policies mandate third-grade retention see significant and persistent increases in high-stakes reading scores in all cohorts. The magnitude of these estimates is similar to that of the "any early literacy policy" estimates described in Section 4.1.1 above, suggesting that states with retention components essentially explain all the average effects of early literacy policies on high-stakes reading scores. By contrast, there is no consistent evidence that high-stakes reading scores increase in states without a retention component.

Impacts on Low-Stakes Reading Scores

Next, we examine the effects of early literacy policies on low-stakes reading scores. Figure 6 presents event-study estimates of early literacy policies' effects on state-level average fourth- and eighth-grade reading NAEP scale scores. We use NAEP scores from 2001 to 2019. During this period, NAEP was administered every two years. For this reason, we create two-year bins of years relative to treatment in the NAEP event-study analysis, so our results cover seven to eight years pre-policy to eight to nine years post-policy.

As with the high-stakes outcomes in SEDA, we find no evidence of statistically significant pre-policy trends in either fourth- or eighth-grade reading. However, *unlike* the high-stakes test outcomes, we find little evidence of changes in reading scores *after* the passage of an early literacy policy. Our estimates are insignificant in both the fourth and eighth grades and across years. The maximum of the 95% confidence intervals would imply an approximately four-point increase in fourth- and eighth-

grade reading. The national standard deviation is roughly 36 scale score points. Thus, these estimates indicate a 0.11 standard deviation increase in NAEP scores. While this is an economically meaningful effect size and is around the upper bound of the 95% confidence interval of the effect on high-stakes SEDA reading scores (Figure 3), the estimates are not statistically significant at any conventional level.

Figure 7 examines the effects of early literacy policies on fourth-grade reading NAEP scale scores in states with comprehensive early literacy policies (Panel A) and non-comprehensive policies (Panel B). Here, we see significant increases in states with comprehensive policies. Similar to the high-stakes reading results, the impacts of comprehensive early literacy policies on fourth-grade NAEP scores phase in as cohorts age into fourth grade until we see a statistically significant, roughly five scale-score point increase six to seven years after the passage of the policy. We find no effect on low-stakes reading scores in states without comprehensive policies. We note some evidence of marginally significant pre-policy decreases in fourth-grade NAEP reading in states with comprehensive early literacy policies relative to never-treated states. While these pre-trend coefficients are only marginally significant and several years before a policy's passage, we interpret these results cautiously. The results for the eighth grade are shown in Appendix Figure A1 and similarly show statistically significant increases in low-stakes reading scores associated with a comprehensive policy. The timing of these eighth-grade impacts is roughly consistent with when fourth-grade students positively affected by comprehensive early literacy policies age into eighth grade. Overall, our estimates suggest that states with comprehensive early literacy policies experience larger high- and low-stakes test score gains than those with non-comprehensive policies.

Figure 8 compares the effects of early literacy policies with third-grade retention requirements (Panel A) with those without (Panel B) on low-stakes NAEP fourth-grade reading assessments.[7] We find no consistent evidence of statistically significant increases in low-stakes tests in states with or without retention requirements. Overall, our results provide evidence that early literacy policies with retention requirements improve high-stakes reading scores more than those without them, but these gains may not translate into increases in low-stakes test scores.

Race to the Top Funding, Early Literacy Policies, and Reading Achievement

As noted above, the introduction of RTTT was contemporaneous with early literacy policies in some states and might confound the effects of early literacy policies. Therefore, we next examine whether the receipt of RTTT funding explains our results. Figure 9 compares the effect of early literacy policies on high-stakes reading achievement separately for states that did not receive RTTT funding (Panel A) and those that did (Panel B). States that did not receive RTTT funding still experienced significant increases in high-stakes reading scores following the passage of an early literacy policy of a similar magnitude as the overall results presented in Panel B of Figure 3. This suggests that RTTT does not drive the effects of early literacy policies on

high-stakes reading achievement. Turning to the states that did receive RTTT funding, they experience effects that are sometimes of a greater magnitude than those of the non-RTTT states, but these estimates are imprecise.[8]

We also compare comprehensive and non-comprehensive policies and policies with and without third-grade retention requirements in states that did not receive RTTT funding (Appendix Figures A3 and A4, respectively). In both cases, the estimates parallel the overall results in Figures 4 and 5, suggesting that RTTT is not driving the effects of early literacy policies on high-stakes reading achievement.[9]

Impacts on High-Stakes Math Test Scores

We repeat our analyses above to determine whether there are any spillover effects of early literacy policies on high-stakes math achievement. Figure 10 shows cohort event-study estimates of the ATT of early literacy policies on high-stakes math scores and parallels Figure 3. The estimates in Panel A are similar to the high-stakes reading achievement results, indicating that early literacy policies do not significantly affect high-stakes math scores for cohorts of students not exposed to the policy. On the other hand, the estimates in Panel B are greater in magnitude than the high-stakes reading estimates, indicating the substantial effect of early literacy policies on high-stakes math scores for cohorts exposed to these policies.

The Impact of Receiving Race to the Top Funding on Math Scores

Next, we examine whether receiving RTTT funding explains the effects of early literacy policies on high-stakes math scores, as shown in Figure 11. Panel A shows that states that do not receive RTTT funding experience no effect on high-stakes math scores, while Panel B shows that the effects are substantial for states receiving RTTT funding. This suggests that RTTT, instead of early literacy policies, may drive math-score effects. As such, we only present subsequent math results for states that did not receive RTTT funding.

Figure 12 compares non-RTTT states with comprehensive early literacy policies to those without comprehensive policies. Similar to the reading results, only states with comprehensive policies appear to experience positive effects on high-stakes math achievement, although the math effects are greater in magnitude than the reading effects. Figure 13 shows the effects for states with and without third-grade retention requirements. Together, these results suggest that third-grade retention mandates drive many of the effects of early literacy policies on high-stakes math achievement, but comprehensive early literacy policies have stronger effects beyond requiring retention alone.

Impacts on Low-Stakes Math Scores

Figure 14 presents event-study estimates of the effects of early literacy policies on state-level average fourth- and eighth-grade math NAEP scale scores. Again, we focus on states that did not receive RTTT funding. Similar to NAEP reading scores, we find

little evidence of changes in math scores after the passage of an early literacy policy. The maximum 95% confidence intervals would imply a five scale-score point increase in eighth-grade math performance and a four-point increase in fourth grade, representing a 0.13 and 0.11 standard deviation increase. Again, the estimates are not statistically significant, except at four to five years post-passage.

Figure 15 examines the effects of early literacy policies on fourth-grade math NAEP scores in states with comprehensive early literacy policies (Panel A) and non-comprehensive policies (Panel B). We see suggestive evidence of an increase in fourth-grade math scores associated with a comprehensive policy but no statistically significant changes in states with noncomprehensive policies. We note some evidence of marginally significant pre-policy decreases in fourth-grade NAEP math scores in states with comprehensive early literacy policies relative to never-treated states. While these pre-trend coefficients are only marginally significant and several years before a policy's passage, we interpret these results cautiously. When we compare the effects of early literacy policies with retention requirements on low-stakes fourth-grade NAEP math assessments in Figure 16, we find no consistent evidence of statistically significant increases in low-stakes tests in states with or without these requirements.

Impacts of High-Stakes Socioeconomic Achievement Gaps

Thus far, the results have shown a substantial increase in high-stakes reading scores for cohorts of students exposed to an early literacy policy, as well as increases in low-stakes scores in states with third-grade retention requirements and comprehensive policies. However, these estimates consider only effects on average test scores and ignore treatment effect heterogeneity. We next examine high-stakes reading test score gaps between non-economically disadvantaged and economically disadvantaged students and between White and Black students.

Figure 17, Panel A shows non-economically disadvantaged and economically disadvantaged test-score gaps on high-stakes reading assessments for cohorts of students exposed to an early literacy policy. Across cohorts, we find no systematic evidence of statistically significant increases in this gap following the introduction of an early literacy policy. Panel B shows White-Black achievement gaps and again finds little evidence of significant changes. In most cases, the coefficient estimates in both panels are generally negative, although not statistically significant. Altogether, we find no systematic evidence that early literacy policies increase achievement gaps and some suggestive evidence that achievement gaps might shrink following the introduction of any early literacy policy.

We also examine how comprehensive early literacy policies (Appendix Figures A8 and A9) and early literacy policies requiring third-grade retention (Appendix Figures A10 and A11) impact socioeconomic and racial test-score gaps and again find no substantial evidence that these types of policies increase the non-economically disadvantaged-economically disadvantaged or White-Black gaps in high-stakes

reading scores. In fact, the post-policy event-study estimates are generally negative and sometimes indicate a statistically significant decrease in test-score gaps, implying that these gaps are diminishing in some cases.

ADDITIONAL ROBUSTNESS CHECKS

Falsification Tests

Our event-study estimates provide no evidence of systematic differential pre-policy trends in high- or low-stakes reading or math scores across treated and never-treated states, supporting the parallel trends assumption. In this section, we provide further evidence that early literacy policies drive our findings by reporting the results of a series of falsification tests wherein we estimate the impact of early literacy policies on state average demographic and macroeconomic conditions. We examine the proportion of a state's non-White population, the proportion living in urban areas, the proportion with a bachelor's degree or higher, and the log median total household income. We use these variables as outcomes in our event-study model. While early literacy policies could impact earnings and educational attainment in the long run, we do not expect them to have any meaningful impact on these state characteristics during our analysis period. Any significant estimates would suggest that differential trends in these characteristics across treated and never-treated states might explain the changes in test scores that we attributed to early literacy policies.

Appendix Figure A12, Panels A through D, shows the results of the falsification tests. We find little evidence of significant differential changes in any of these demographic or macroeconomic conditions before or after the introduction of an early literacy policy, except for small and marginally significant increases in the percentage of non-white individuals in states with early literacy policies five years post-policy adoption. These results provide additional support for the interpretation of our results as the causal effects of early literacy policies.

Permutation Tests

Next, we examine whether the documented impacts of early literacy policies on student achievement could result from random chance. We examine the likelihood that we would detect treatment effect estimates of a similar magnitude by randomly assigning treatment status 500 times, keeping the time distribution of adoption constant (i.e., the same number of states adopt an early literacy policy each year as were adopted in reality). We then estimate the event-study model and collect the coefficient estimates. We construct empirical p-values by computing the proportion of placebo treatment effects greater than the actual treatment effect and empirical confidence intervals by calculating the 5^{th} and 95^{th} percentiles of the placebo treatment effect distribution.

Appendix Figure A13 presents the results of this exercise for high-stakes third- to eighth-grade SEDA reading scores. The solid lines represent our event-study estimates with states' actual treatment status, as shown in Figure 3. The shaded areas represent empirical 95% confidence intervals. Event-study estimates outside those confidence intervals are larger in magnitude than the 95th or 5th percentile of placebo event-study estimates, suggesting a statistically significant estimate. Our results are robust to these empirical confidence intervals. In the years leading up to the adoption of an early literacy policy, we see no evidence of significant differential changes in high-stakes reading scores, except in eighth grade. Immediately following the introduction of an early literacy policy, we see that our estimated increases in reading scores are significantly beyond what one might expect if treatment were randomly assigned in third through fifth grade but still inside the empirical confidence intervals in sixth through eighth grade, consistent with our prior findings.

Missing Data Analysis

Another concern regarding the validity of our analyses is differential attrition from the data. We have complete NAEP data for overall average test scores throughout the analysis period. However, some states did not report their high-stakes standardized test scores in specific years, and these outcomes are missing from the SEDA data. If the probability that a state is missing test score data in SEDA correlates with having an early literacy policy, our estimates may be biased. This bias would be particularly concerning if states were less likely to report their scores in years of poor performance. This attrition would bias our estimates upward, potentially leading us to conclude that early literacy policies improve achievement when no such effects exist.

We test for differential attrition by estimating our preferred event-study model with an indicator for missing test score data in a given year. Appendix Figure A14 shows the results of this exercise for missing SEDA reading scores for third through eighth grades. We find no evidence that states adopting early literacy policies were more or less likely to have missing test score information than never-treated states, suggesting that differential attrition is not driving our findings.

Alternative Estimators

Finally, we test the robustness of our main estimates using the Callaway and Sant'Anna (2021) estimator to two alternative difference-in-differences estimators: (1) two-way fixed-effects using ordinary least squares (OLS) and (2) the interaction-weighted event-study estimator proposed by Sun and Abraham (2021). Both estimation methods model the event study as follows:

$$Y_{st} = \alpha + \sum_{j=[-5,5]} \delta_j \, \mathbf{1}(t - E_s = j) + \theta_s + \tau_t + \epsilon_{st} \qquad (3)$$

Again, the outcome, Y_{st}, is the average test score in state *s* in year *t*. E_s is the year in which state *s* first implemented an early literacy policy. $\mathbf{1}(t - E_s = j)$ are indicators

that equal one when the time relative to treatment ($t - E_s$**)** equals *j.* We omit *t-1* as the reference year. θ_s are state fixed effects and τ_t are year fixed effects. ϵ_{st} are state-year-specific idiosyncratic errors. The primary coefficients of interest are δ_j, which measure how different average test scores are *j* years after (or before for negative values of *j*) the introduction of an early literacy policy relative to states without policies in the same period.

First, we estimate the model using OLS. Recent econometric research has identified issues with this method in the presence of staggered adoption and treatment effect heterogeneity (e.g., Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Roth et al., 2022; Sun & Abraham, 2021). In particular, the event-study parameters represent weighted averages of all possible 2x2 difference-in-differences, where the weights may be negative. However, the extent to which this is an issue is unclear a priori. The OLS estimates of Equation (3) for the SEDA reading scores are presented in Appendix A15. These results are consistent with our preferred estimation strategy.

Next, we estimate Equation (3) using the interaction-weighted event-study estimator from Sun and Abraham (2021). This method begins by estimating Equation (3) using OLS and then reweights the event-study parameters such that the weights are non-negative. These estimates are presented in Appendix Figure A16. Again, we find broadly similar results to our preferred estimation method. This makes sense given that the Sun and Abraham (2021) estimator is simply a special case and aggregation of the Callaway and Sant'Anna (2021) estimator in models without covariates. Altogether, these exercises indicate that our choice of estimator does not drive the findings.

DISCUSSION AND CONCLUSION

In this study, we investigate the causal effects of early literacy policies on high- and low-stakes reading and math test scores using an event-study research design that leverages differences in the adoption and content of early literacy policies across states and over time. Our findings contribute to the literature on early literacy policies in several ways. First, we examine their impact on both high- and low-stakes assessments, allowing us to differentiate between test-taking and human capital improvements. Second, we leverage cross-state variation in the content of early literacy policies to examine the effects of specific components. Finally, our national setting provides external validity for our findings.

Our results provide compelling evidence that early literacy policies improve high-stakes achievement in the short term but that the composition of these policies matters. Having any early literacy policy improves high-stakes reading scores in the elementary years for cohorts of students exposed to the policy. We also find substantial increases in high-stakes math performance following the introduction of an early literacy policy, but further analysis suggests that these gains may be due to other education policy changes such as RTTT.

We find little evidence of significant increases in low-stakes reading test scores, except in states with the most comprehensive policies. We also find that the largest impacts on high-stakes test scores are in states with comprehensive early literacy policies and third-grade retention requirements. Furthermore, the high-stakes test-score gains appear to be distributed equitably, potentially decreasing socioeconomic and racial test-score gaps.

Taken together, our results suggest that noncomprehensive early literacy policies provide potentially superficial gains in reading and highlight the importance of using low-stakes test scores to evaluate the impact of education policies. Focusing on high-stakes test scores can mislead policymakers if there are policy-induced changes in high-stakes test-taking or tested materials that do not reflect changes in human capital. Examining low-stakes tests can provide a better measure of changes in actual learning. In addition, our findings underscore the importance of the content and incentives of early literacy policies. The best evidence for significant increases in both low- and high-stakes test scores comes from the states with the most comprehensive early literacy policies, including those with third-grade retention requirements. Altogether, these results indicate that the full set of interventions available under early literacy policies is important in improving literacy achievement and skills.

Although our study sheds light on the potential benefits of early literacy policies, there are some limitations that point to areas for future research. For example, while we provide evidence that comprehensive early literacy policies and retention mandates play an important role in improving state summative assessment scores, we cannot examine the mechanisms by which these policy components improve outcomes. Further research on the implementation of these policy components is therefore vital to understanding how early literacy policies operate. Additionally, we only focus on short-run test-score outcomes. However, prior work has established the importance of early literacy skills in determining non-cognitive outcomes and long-term student success (Cunningham & Stanovich, 1997; Fiester & Smith, 2010; Hernandez, 2011; Sparks et al., 2014). To fully understand the benefits of early literacy policies, it is important to enumerate their non-cognitive and long-term impacts. Finally, this study does not examine the costs associated with early literacy policies. While we show substantial short-run high-stakes test score gains, policymakers must weigh all the benefits against the costs of early literacy policies in the short and long run.

ENDNOTES

[1] The NAEP is administered approximately every two years. While 2021 is the most recent year available, we exclude it due to COVID-19 pandemic-related disruptions to schooling.
[2] Before 2009 for the SEDA data and before 2003 for the NAEP data.
[3] We examine the possibility that missing test-score data correlates with early literacy policies in Section 5.4.
[4] The full table, including all grade levels, is available in Appendix Table A1.
[5] We estimate models with pre-treatment covariates but find no substantial improvements in efficiency. Therefore, we prefer the parsimonious specification without controls.
[6] Because the NAEP is administered every two years, we bin the event-study estimates into two-year bins such that we have five bins before treatment (starting 7 to 8 years pre-policy) and five bins after treatment (ending 8 to 9 years post-policy).
[7] Parallel results for 8th grade NAEP reading scores are in Appendix Figure A2.
[8] Appendix Figure A5 shows the low-stakes reading score results are also robust to omitting states that received RTTT funding.
[9] Appendix Figures A6 and 7 show the low-stakes reading results are similarly robust.

REFERENCES

Burns, J. (2016, March 10). Should struggling readers be retained in 3rd grade? [Michigan State University College of Education]. *Green & Write*. https://education.msu.edu/green-and-write/2016/should-struggling-readers-be-retained-in-3rd-grade/

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*, 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Common Core State Standards Initiative. (n.d.). *About the standards*. Common Core State Standards Initiative. Retrieved July 11, 2022, from http://www.corestandards.org/about-the-standards/

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945.

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, *39*(6), 507–514. https://doi.org/10.1177/00222194060390060301

Duke, N. K., Moje, E. B., & Palincsar, A. S. (2014). *Three IRA Literacy Research Panel Members Comment on Third Grade Retention Laws*. *47*(1), 6.

Eren, O., Lovenheim, M. F., & Mocan, H. N. (2022). The effect of grade retention on adult crime: Evidence from a test-based promotion policy. *Journal of Labor Economics*, *40*(2).

ExcelinEd. (2021). *Comprehensive early literacy policy: State-by-state analysis of fundamental principles*. ExcelinEd. https://excelined.org/wp-content/uploads/2021/10/ExcelinEd_PolicyToolkit_EarlyLiteracy_StatebyStateAnalysis_2021.pdf

Fiester, L. (2013). *Early warning confirmed: A research update on third-grade reading*. The Annie E. Casey Foundation.

Fiester, L., & Smith, R. (2010). *Early warning! Why reading by the end of third grade matters*. The Annie E. Casey Foundation.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the Black-White Test Score Gap in the First Two Years of School. *The Review of Economics and Statistics*, *86*(2), 447–464. https://doi.org/10.1162/003465304323031049

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2021.03.014

Greene, J. P., & Winters, M. A. (2004). *An evaluation of Florida's program to end social promotion* (No. 7; Education Working Paper). Center for Civic Innovation at the Manhattan Institute.

Greene, J. P., & Winters, M. A. (2006). *Getting farther ahead by staying behind: A second-year evaluation of Florida's policy to end social promotion* (No. 49; Civic Report). Center for Civic Innovation at the Manhattan Institute.

Greene, J. P., & Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, *2*(4), 319–342.

Greene, J. P., & Winters, M. A. (2009). The effects of exemptions to Florida's test-based promotion policy: Who is retained? Who benefits academically? *Economics of Education Review*, *28*, 135–142.

Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation*. The Annie E. Casey Foundation.

Holmes, C. T., & Matthews, K. M. (1984). The effects of non promotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research*, *52*(4), 225–236. https://doi.org/10.3102%2F00346543054002225

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis*, *29*(4), 239–261. https://doi.org/10.3102%2F0162373707309073

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, *86*(1), 226–244.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, *30*(3), 420–437.

Licalsi, C., Ozek, U., & Figlio, D. (2019). The uneven implementation of universal school policies: Maternal education and Florida's mandatory grade retention policy. *Education Finance and Policy*, *14*(3), 383–413. https://doi.org/10.1162/edfp_a_00252

Livingston, D. R., & Livingston, S. M. (2002). Failing Georgia: The case against the ban on social promotion. *Education Policy Analysis Archives*, *10*(49).

Lorence, J. (2014). Third-grade retention and reading achievement in Texas: A nine year panel study. *Social Science Research*, *48*, 1–19.

McCombs, J. S., Kirby, S. N., & Mariano, L. T. (2009). *Ending social promotion without leaving children behind: The case of New York City* (1st ed.). RAND Corporation. http://www.jstor.org/stable/10.7249/mg894nycdoe

Nagaoka, J., & Roderick, M. (2004). *Ending social promotion: The effects of retention* (Charting Reform in Chicago Series). Consortium on Chicago School Research.

National Assessment of Educational Progress. (2022). *Reading and mathematics scores decline during COVID-19 pandemic*. National Center for Education Statistics. https://www.nationsreportcard.gov/highlights/ltt/2022/

National Center for Education Statistics. (2020). Reading performance. In *The Condition of Education 2020*. National Center for Education Statistics. https://nces.ed.gov/programs/coe/pdf/coe_cnb.pdf

Oklahoma State Department of Education. (2020). *Reading Sufficiency Act study: Prepared by Oklahoma State Department of Education, 2019-2020 school year*. Oklahoma State Department of Education. https://sde.ok.gov/sites/default/files/documents/files/2020%20Governor%27s%20Report%20FINAL.pdf

Popham, W. J. (2001). Teaching to the test? *Educational Leadership*, *58*(6), 16–20.

Reardon, S. F., Weathers, E., Fahle, E., Jang, H., & Kalogrides, D. (2019). *Is Separate Still Unequal? New Evidence on School Segregation and Racial Academic Achievement Gaps*. https://vtechworks.lib.vt.edu/handle/10919/97804

Reardon, S., Kalogrides, D., Ho, A., Shear, B., Fahle, E., Jang, H., & Chavez, B. (2021). *Stanford Education Data Archive (SEDA)* [English]. Stanford University. https://exhibits.stanford.edu/data/catalog/db586ns4974

Roderick, M., & Nagaoka, Jenny. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, *27*(4), 309–340.

Roth, J., Sant'Anna, P. H. C., Bikinski, A., & Poe, J. (2022). *What's trending in difference-in-differences? A synthesis of the recent econometrics literature*. https://jonathandroth.github.io/assets/files/DiD_Review_Paper.pdf

Sant'Anna, P. H. C., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, *219*(1), 101–122. https://doi.org/10.1016/j.jeconom.2020.06.003

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154–169.

Sparks, R. L., Patton, J., & Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of "10 years later." *Read Writ*, *27*, 189–211.

Starr, A. (2019, July 13). States are ratcheting up reading expectations for 3rd-graders. *National Public Radio*. https://www.npr.org/2019/07/13/741156019/states-are-ratcheting-up-reading-expectations-for-3rd-graders

Strunk, K. O., Wright, T. S., Kilbride, T., Zhu, Q., Cummings, A., West, J., Turner, M., & De Voto, C. (2021). *Michigan's Read by Grade Three Law: Year one report*. Education Policy Innovation Collaborative. https://epicedpolicy.org/wp-content/uploads/2021/03/Year_One_RBG3_Report.pdf

Strunk, K. O., Wright, T. S., Westall, J., Zhu, Q., Kilbride, T., Cummings, A., Utter, A., & Mavrogordato, M. (2022). *Michigan's Read by Grade Three Law: Year two report*. Education Policy Innovation Collaborative. https://epicedpolicy.org/wp-content/uploads/2022/02/RBG3_Rpt_Yr2_Feb2022.pdf

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*, 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

The Nation's Report Card. (n.d.). *NAEP data explorer*. Retrieved January 26, 2022, from https://www.nationsreportcard.gov/ndecore/xplore/NDE

The White House. (2017). *Race to the Top*. The White House, President Barack Obama. https://obamawhitehouse.archives.gov/issues/education/k-12/race-to-the-top

Valencia, R. R., & Villarreal, B. J. (2004). Texas' second wave of high-stakes testing: Anti-social promotion legislation, grade retention, and adverse impact on minorities. In *Leaving children behind: How "Texas-style" accountability fails Latino youth*. SUNY Press.

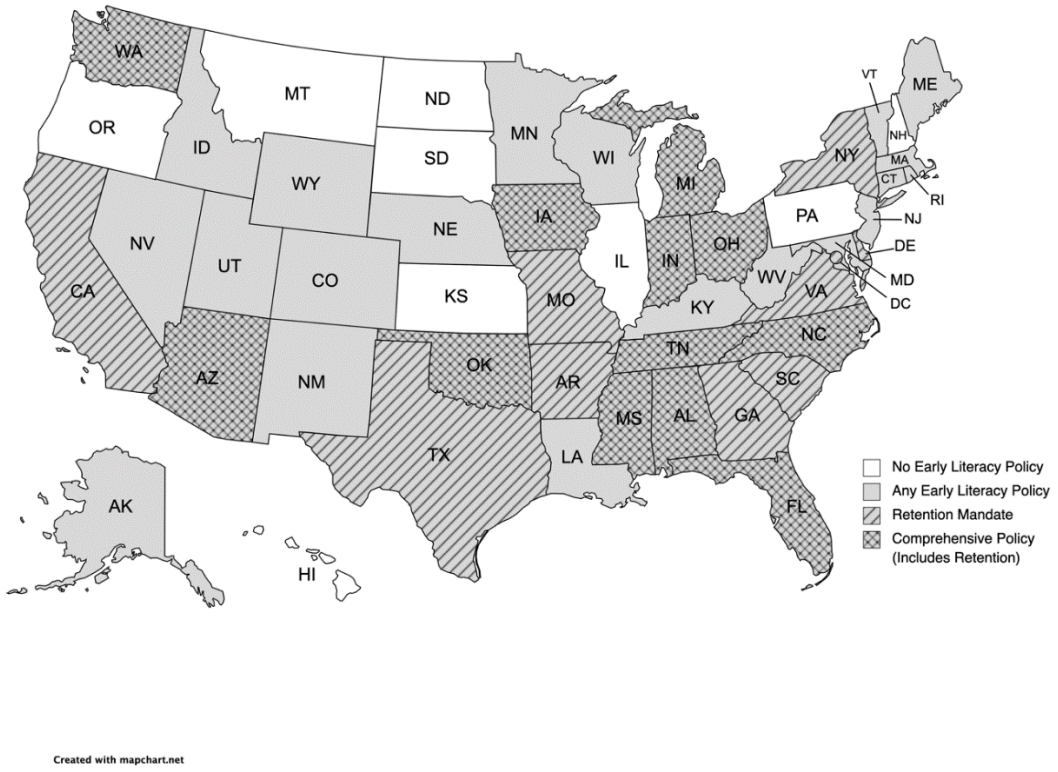
Weiss, S., Stallings, D. T., & Porter, S. (2018). *Is Read to Achieve making the grade? An assessment of North Carolina's elementary reading proficiency initiative*. North Carolina State University College of Education.

Winters, M. A., & Greene, J. P. (2012). The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy*, *7*(3), 305–330.

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educational Psychology*, *102*(1), 135–152. https://psycnet.apa.org/doi/10.1037/a0016664

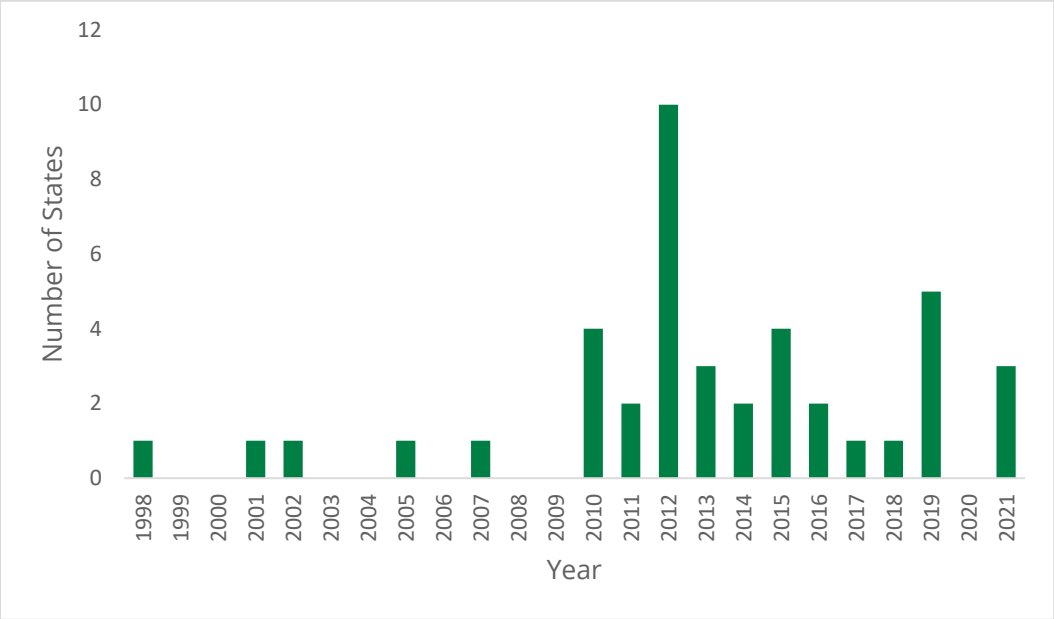
FIGURES AND TABLES

Figure 1. Map of States' Early Literacy Policy Components



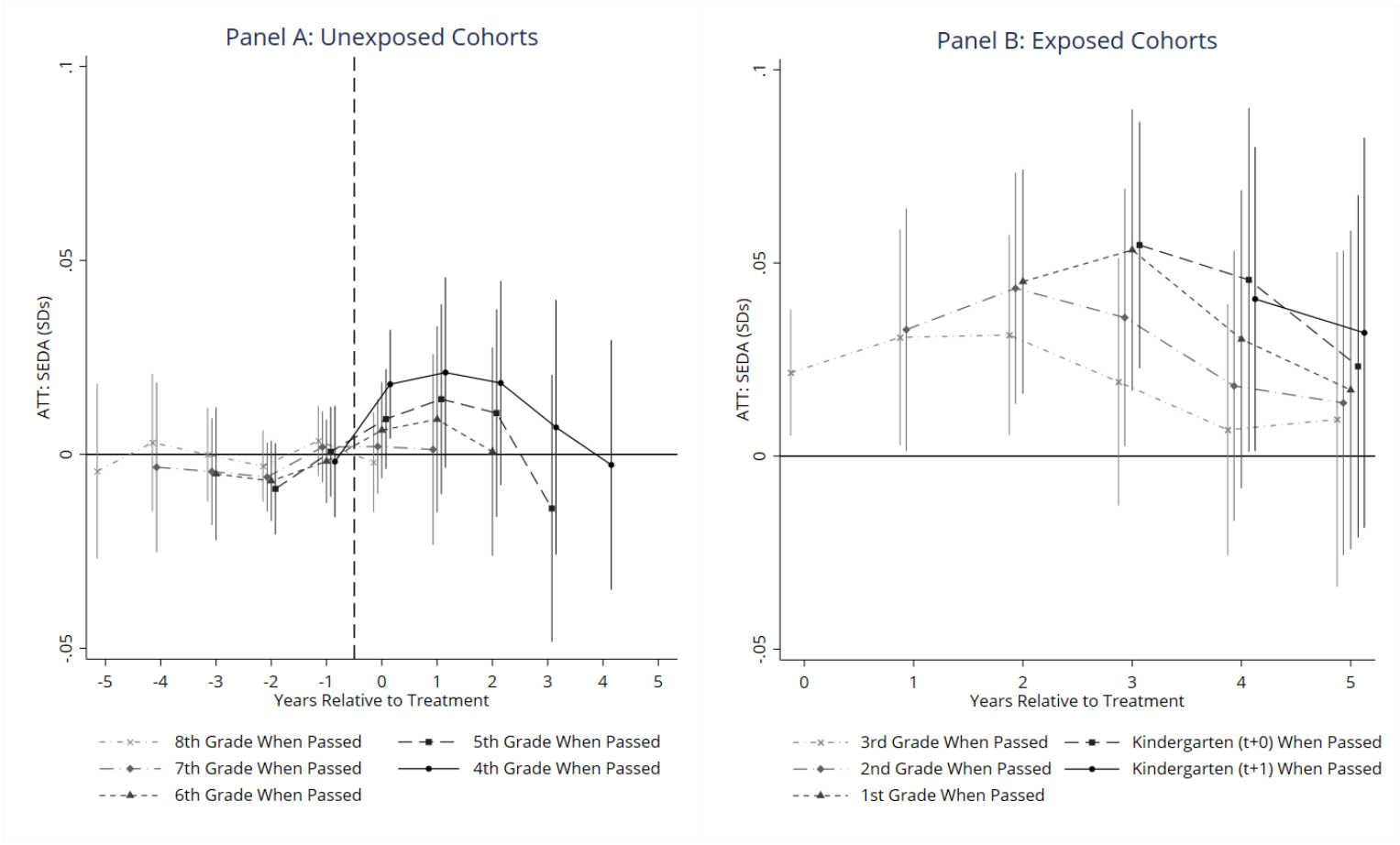
Note: Map is based on data from ExcelinEd (2021).

Figure 2. Number of States Passing Early Literacy Policies, by Year



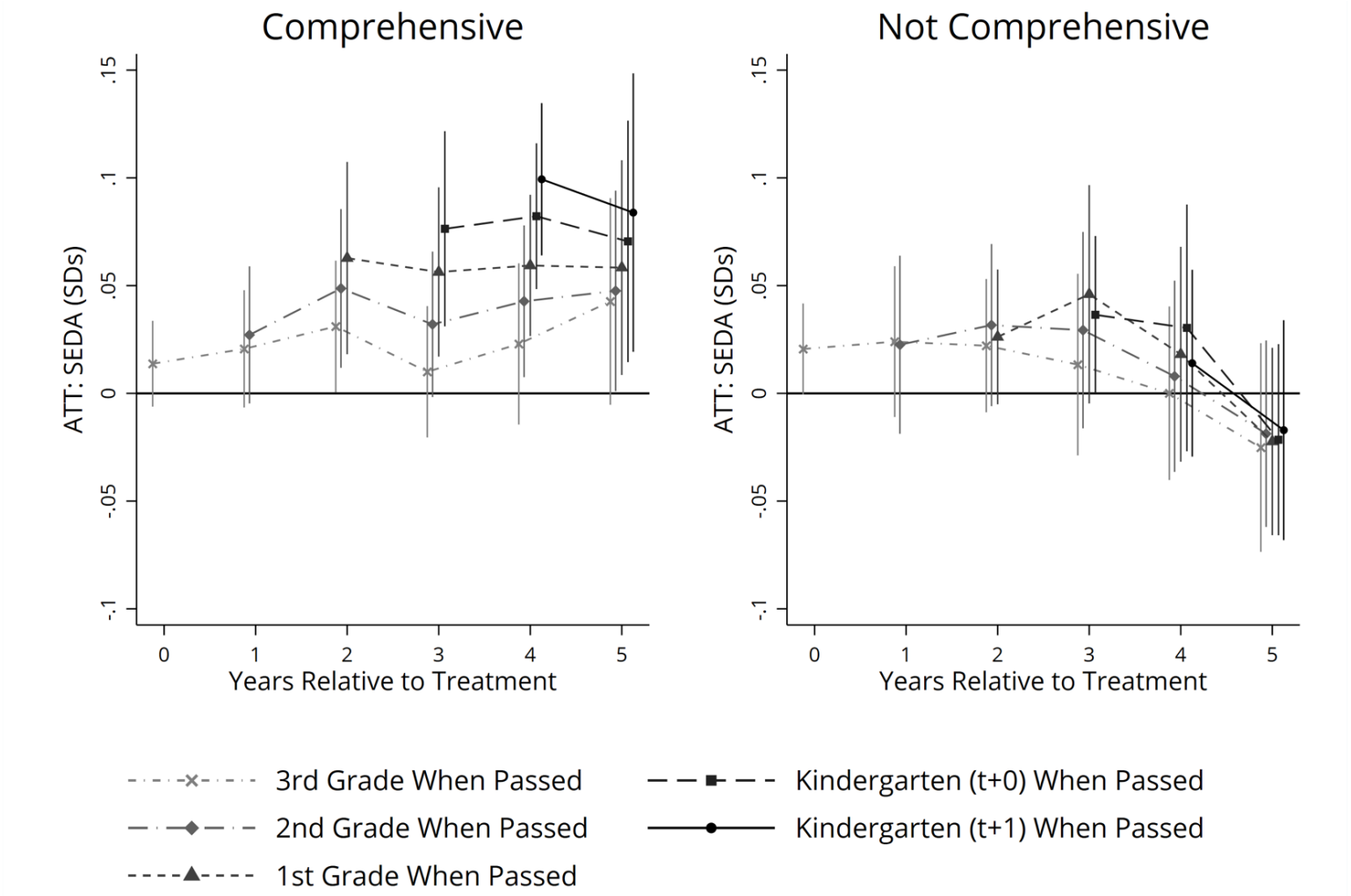
Note. Data derived from ExcelinEd (2021). The figure shows the number of states passing their first early literacy policy component in the year given on the x-axis.

Figure 3. Early Literacy Policies and High-Stakes Reading Scores



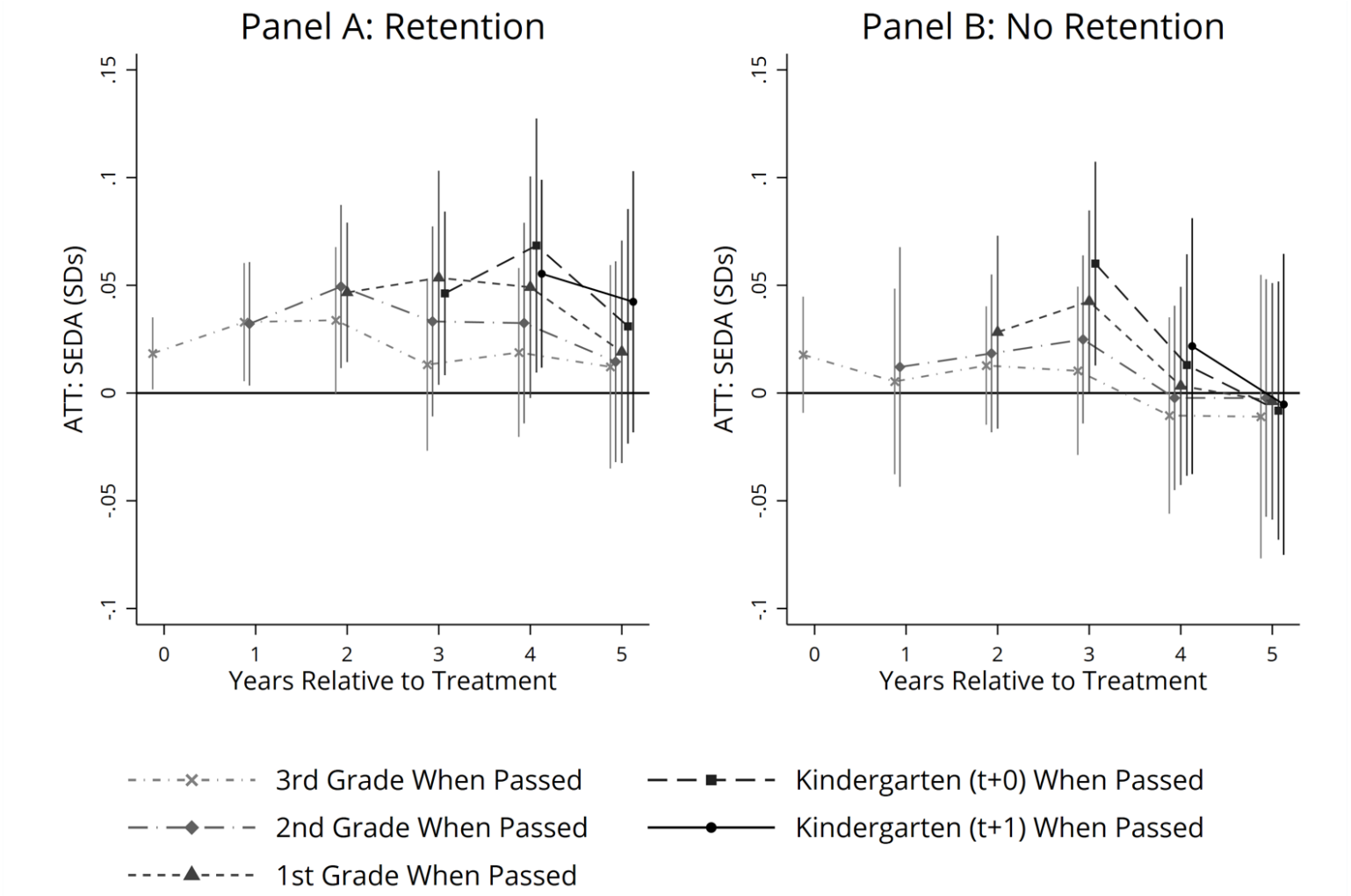
Note: Data are from overall average SEDA reading scores, 2009-2018. Panel A includes cohorts not exposed to early literacy policies in K-3. Panel B includes cohorts exposed to early literacy policies in K-3. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 4. Comprehensive Early Literacy Policies and High-Stakes Reading Achievement



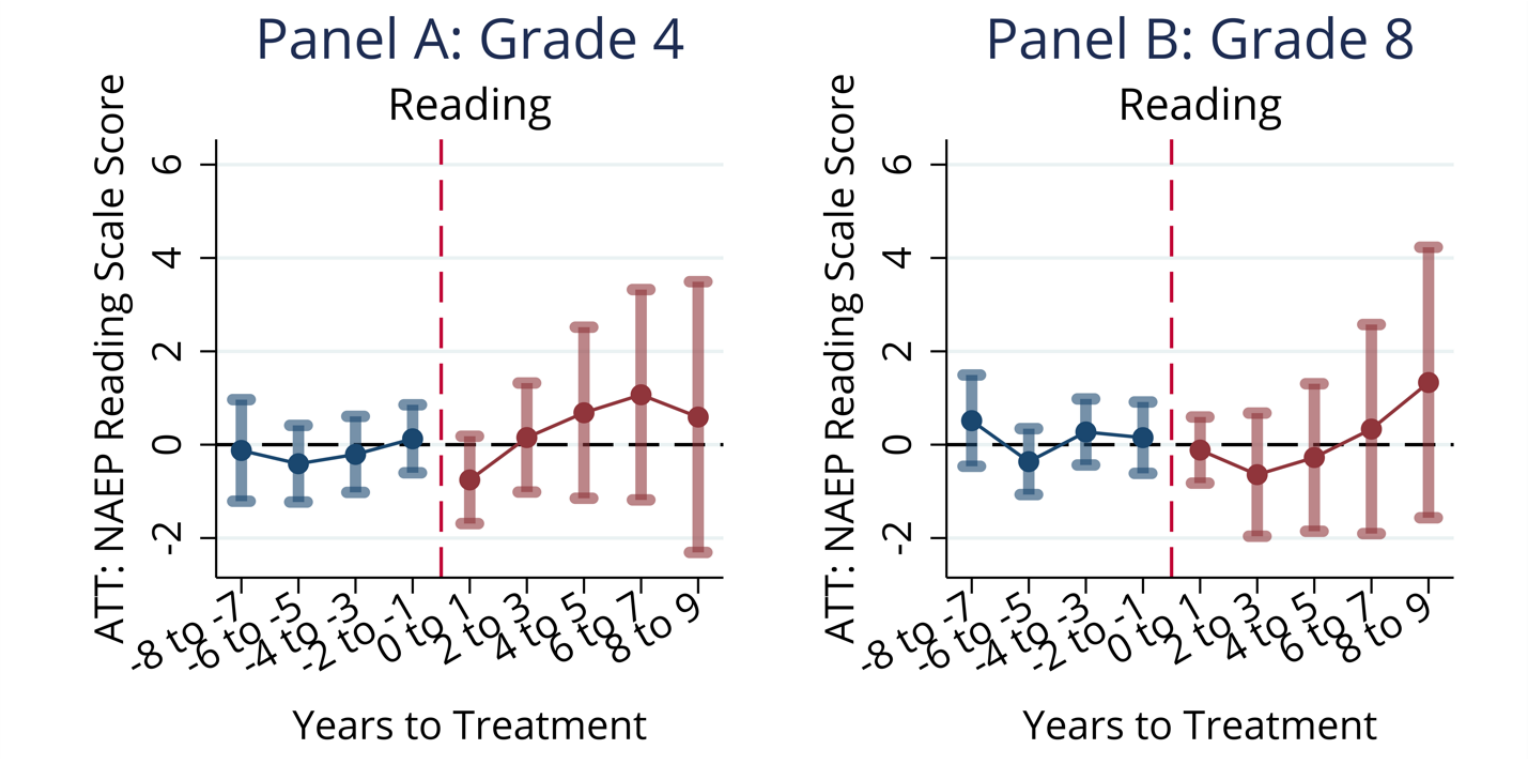
Note: Data are from overall average SEDA reading scores, 2009-2018. Panel A includes states with comprehensive early literacy policies and never-treated states. Panel B includes states without comprehensive early literacy policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 5. Early Literacy Policies with Retention Requirements and High-Stakes Reading Achievement



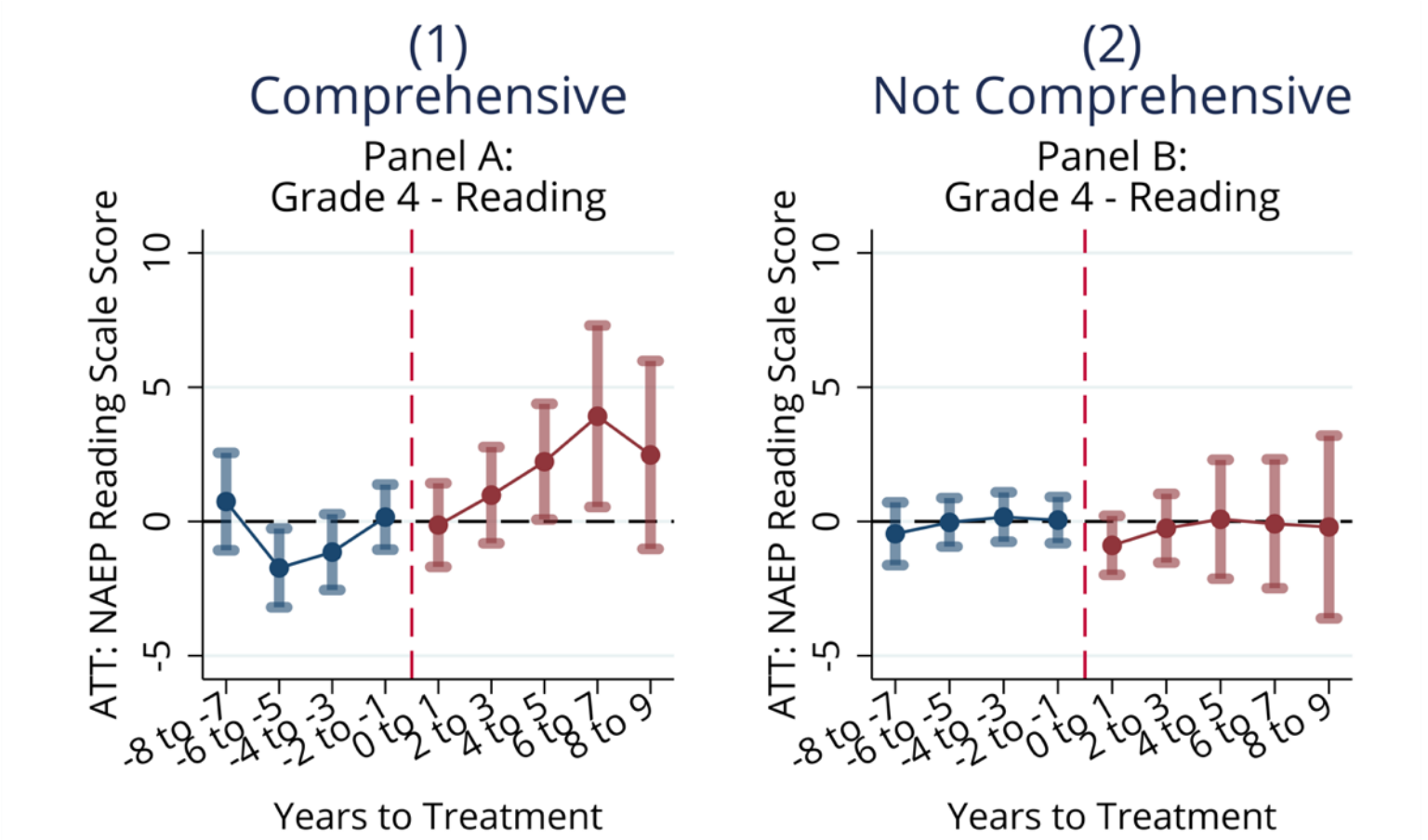
Note: Data are from overall average SEDA reading scores, 2009-2018. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 6. Early Literacy Policies and Low-Stakes Reading Scores



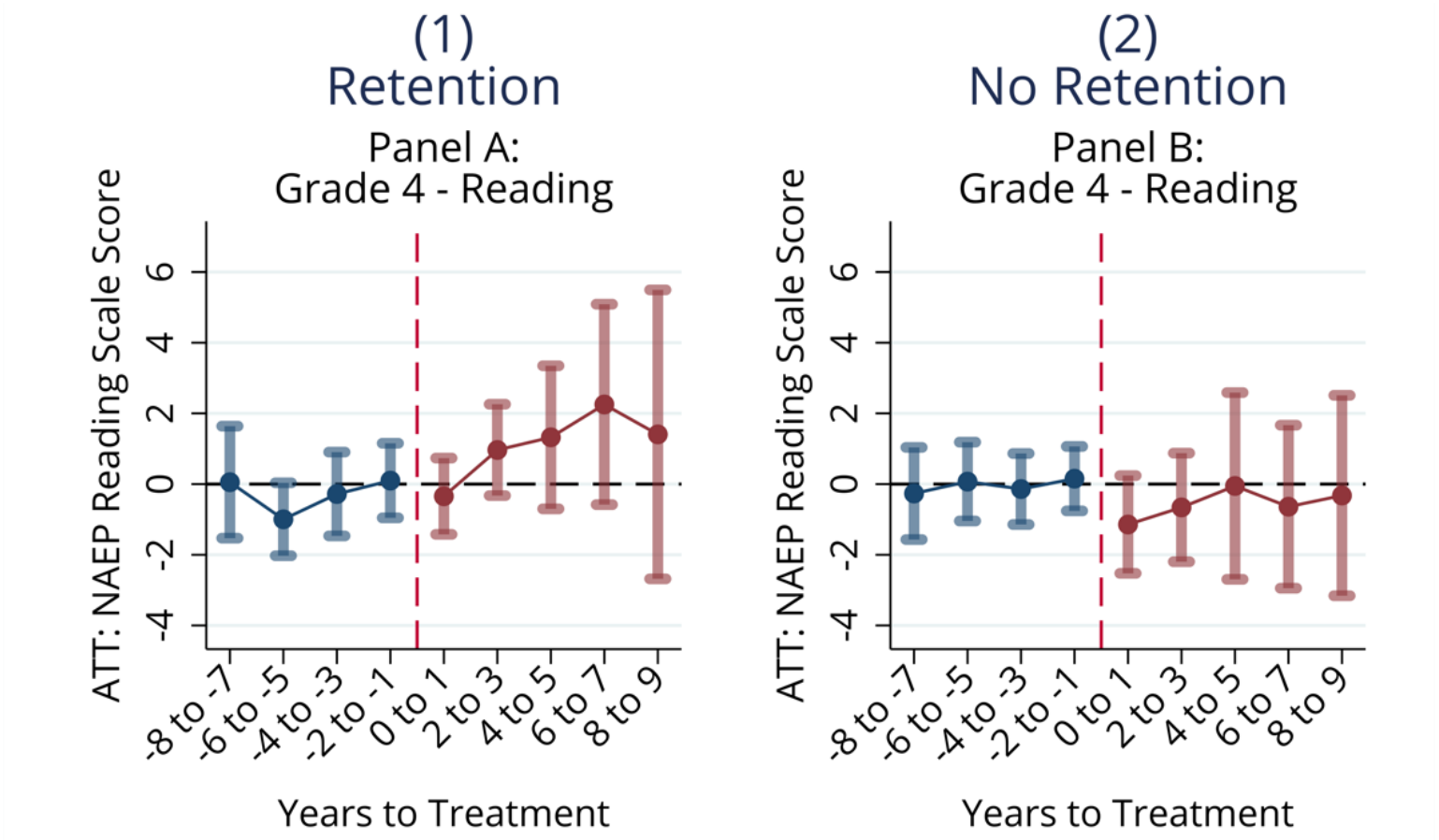
Note: Data are from overall average NAEP fourth- and eighth-grade reading scale scores, 2003-2019. Panel A and B examine fourth- and eighth-grade NAEP reading scores, respectively. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 7. Comprehensive Early Literacy Policies and Low-Stakes Reading Scores



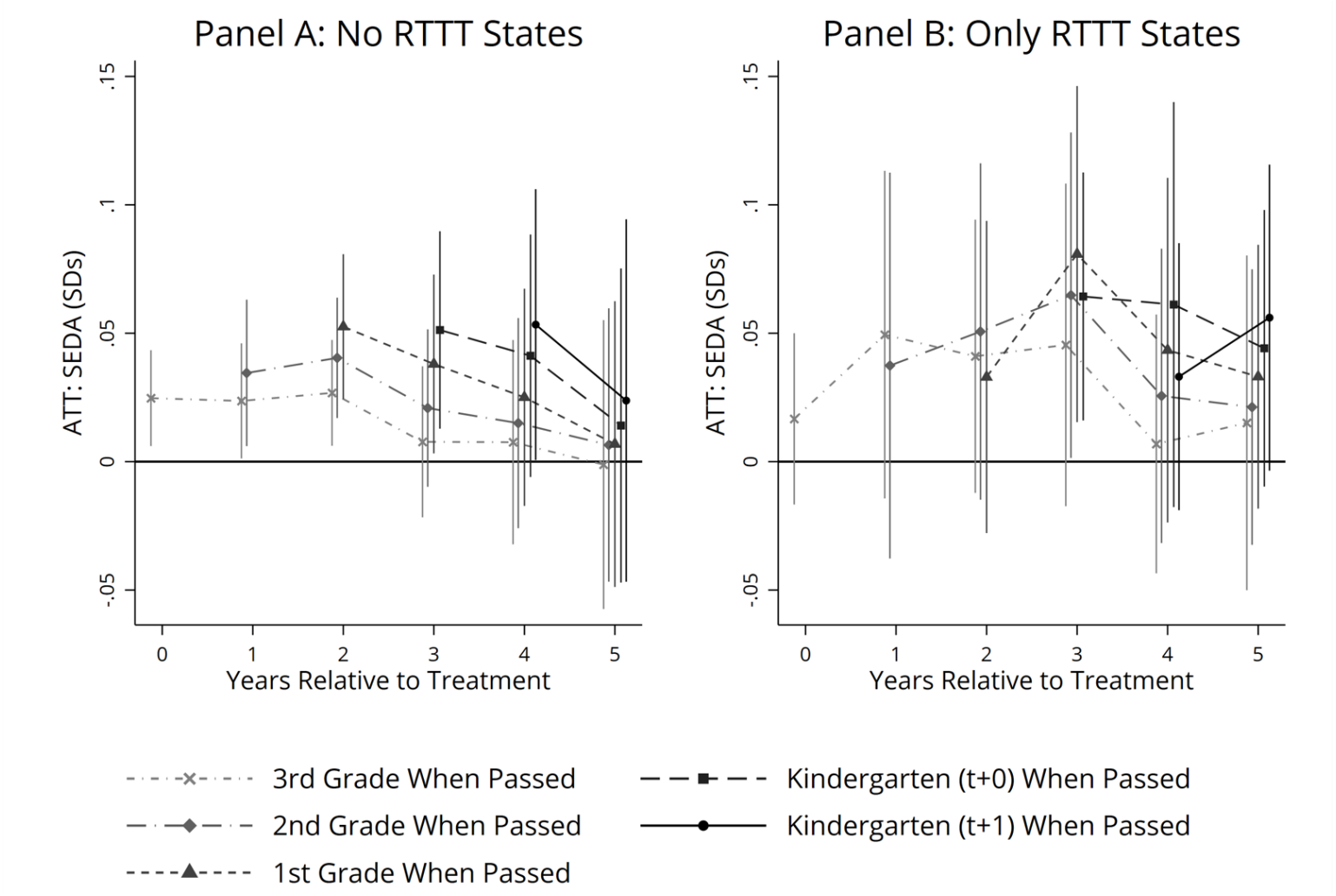
Note: Data are from overall average NAEP fourth-grade reading scale scores, 2003-2019. Column (1) includes states with comprehensive policies and never-treated states. Column (2) includes states with non-comprehensive policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 8. Early Literacy Policies with Retention Requirements and Low-Stakes Reading Scores



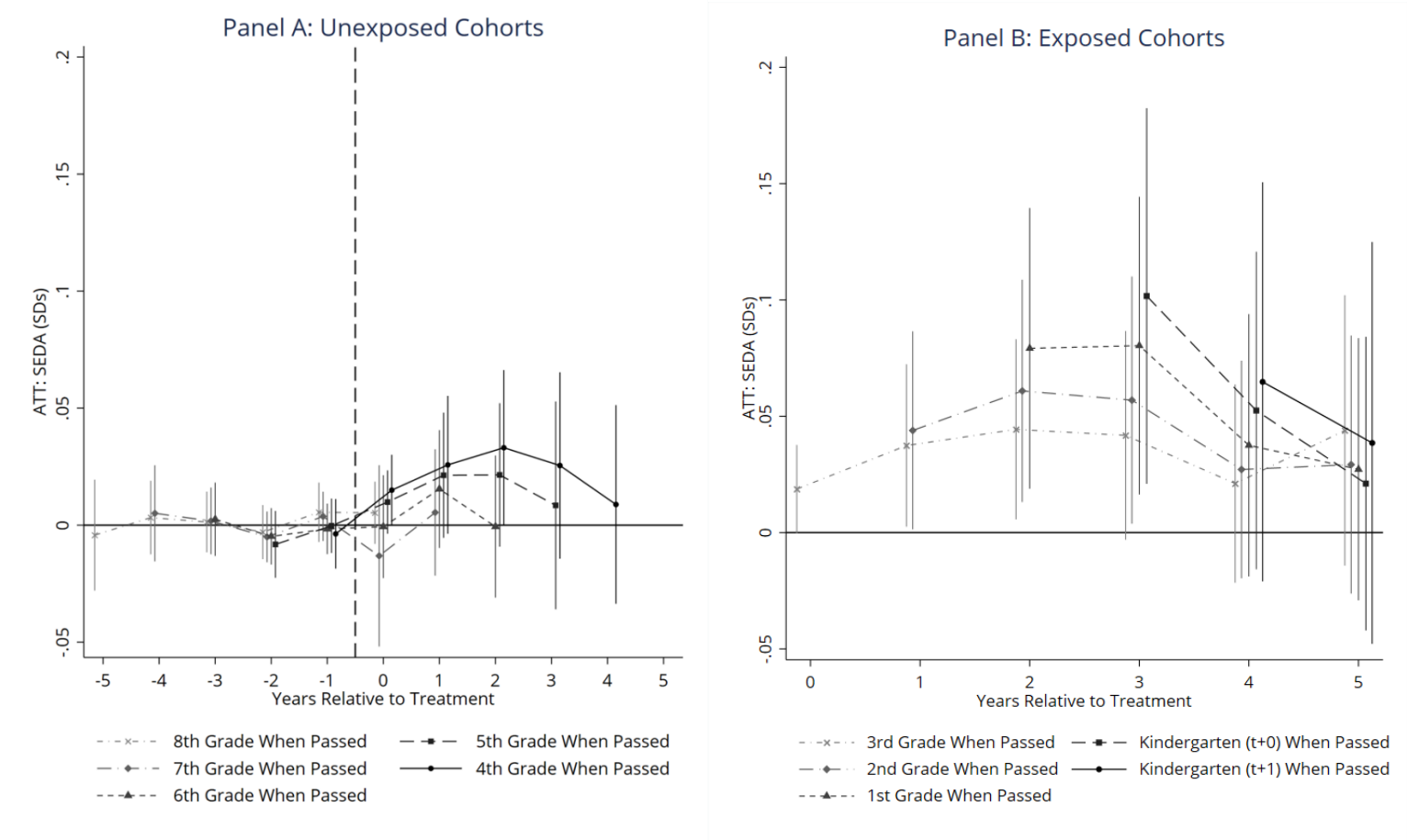
Note: Data are from overall average NAEP fourth-grade reading scale scores, 2003-2019. Column (1) includes states with retention requirements and never-treated states. Column (2) includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 9. Early Literacy Policies, Race to the Top (RTTT), and High-Stakes Reading Scores



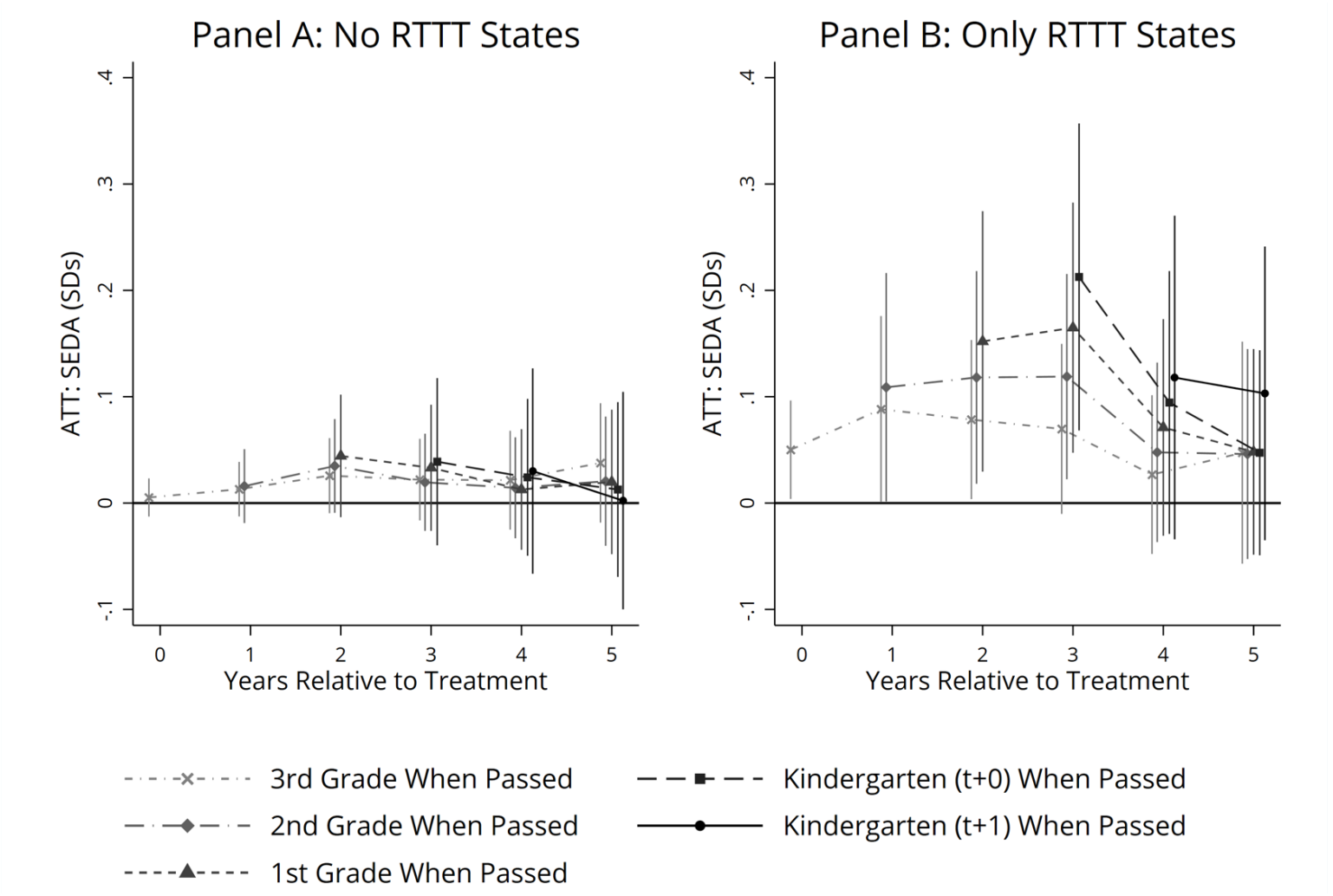
Note: Data are from overall average SEDA reading scores, 2009-2018. Panel A includes states that did not receive Race to the Top funding and never-treated states. Panel B includes states that received Race to the Top funding and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 10. Early Literacy Policies and High-Stakes Math Scores



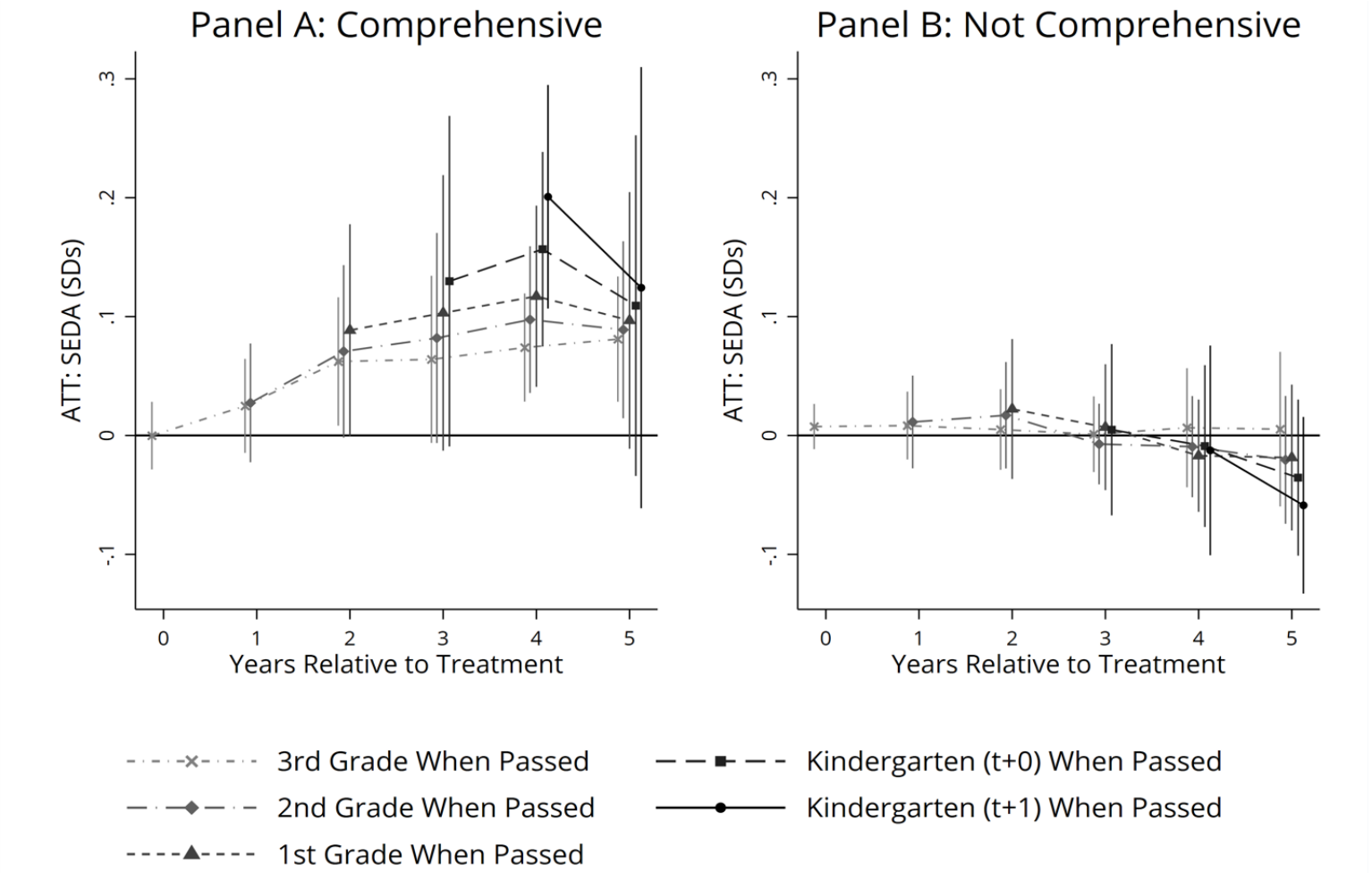
Note: Data are from overall average SEDA math scores, 2009-2018. Panel A includes cohorts not exposed to early literacy policies in K-3. Panel B includes cohorts exposed to early literacy policies in K-3. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 11. Early Literacy Policies, Race to the Top, and High-Stakes Math Scores



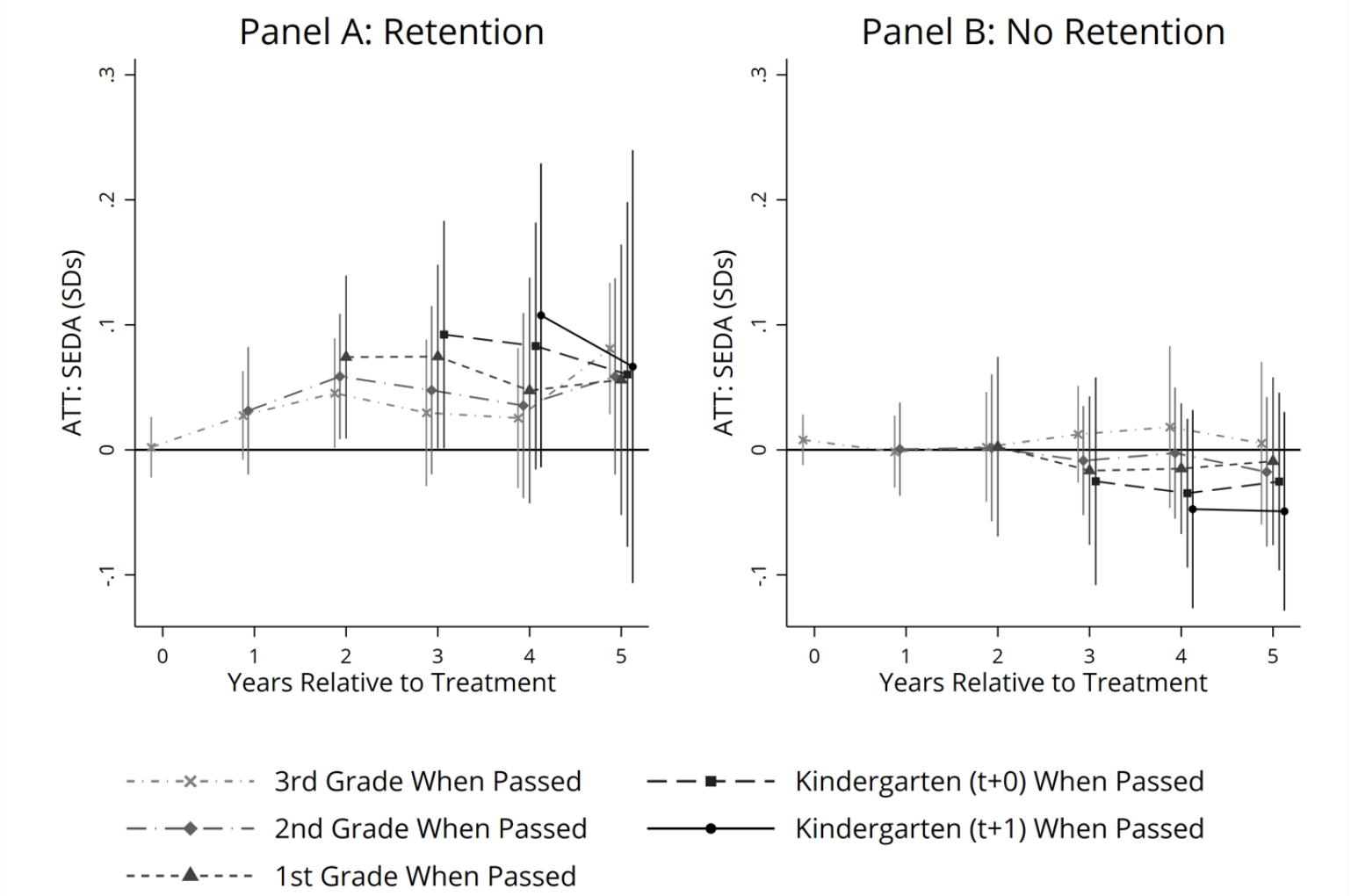
Note: Data are from overall average SEDA math scores, 2009-2018. Panel A includes states that did not receive Race to the Top funding and never-treated states. Panel B includes states that received Race to the Top funding and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 12. Comprehensive Early Literacy Policies, No Race to the Top, and High-Stakes Math Achievement



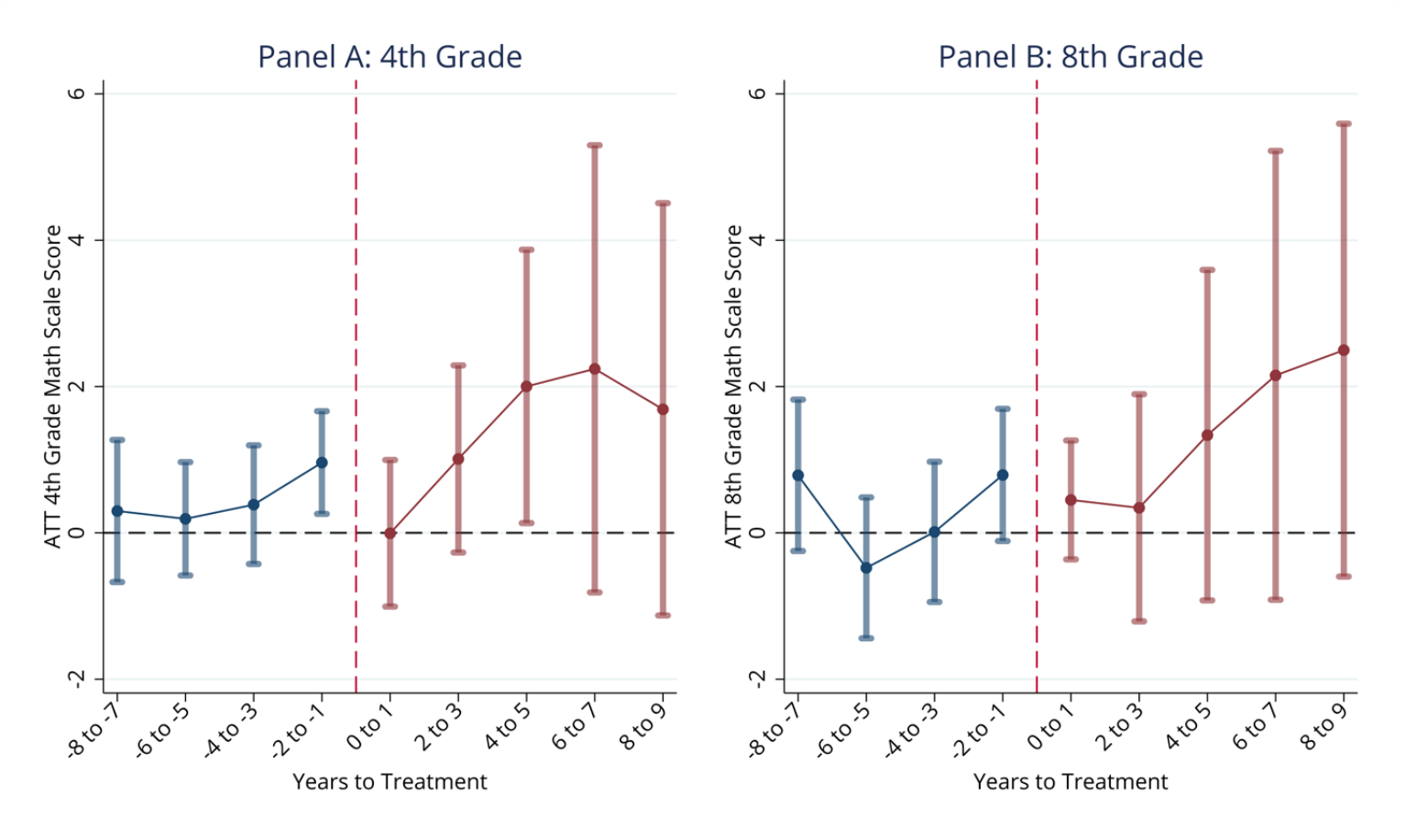
Note: Data are from overall average SEDA math scores, 2009-2018. We include only states that did not receive Race to the Top funding. Panel A includes states with comprehensive early literacy policies and never-treated states. Panel B includes states without comprehensive early literacy policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Figure 13. Early Literacy Policies with Retention Requirements and High-Stakes Math Test Scores



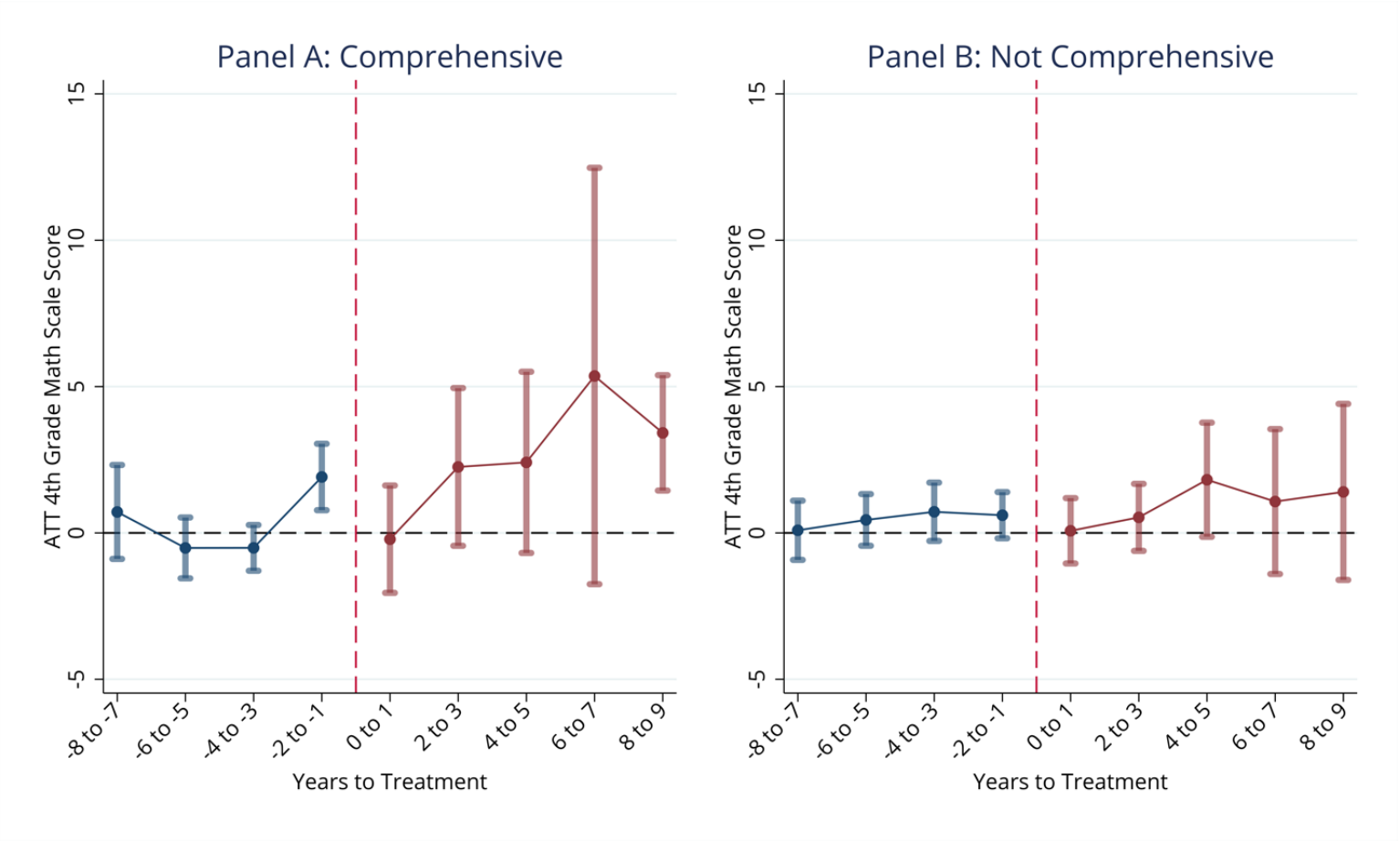
Note: Data are from overall average SEDA math scores, 2009-2018. We include only states that did not receive Race to the Top funding. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level

Figure 14. Early Literacy Policies and Low-Stakes Math Scores



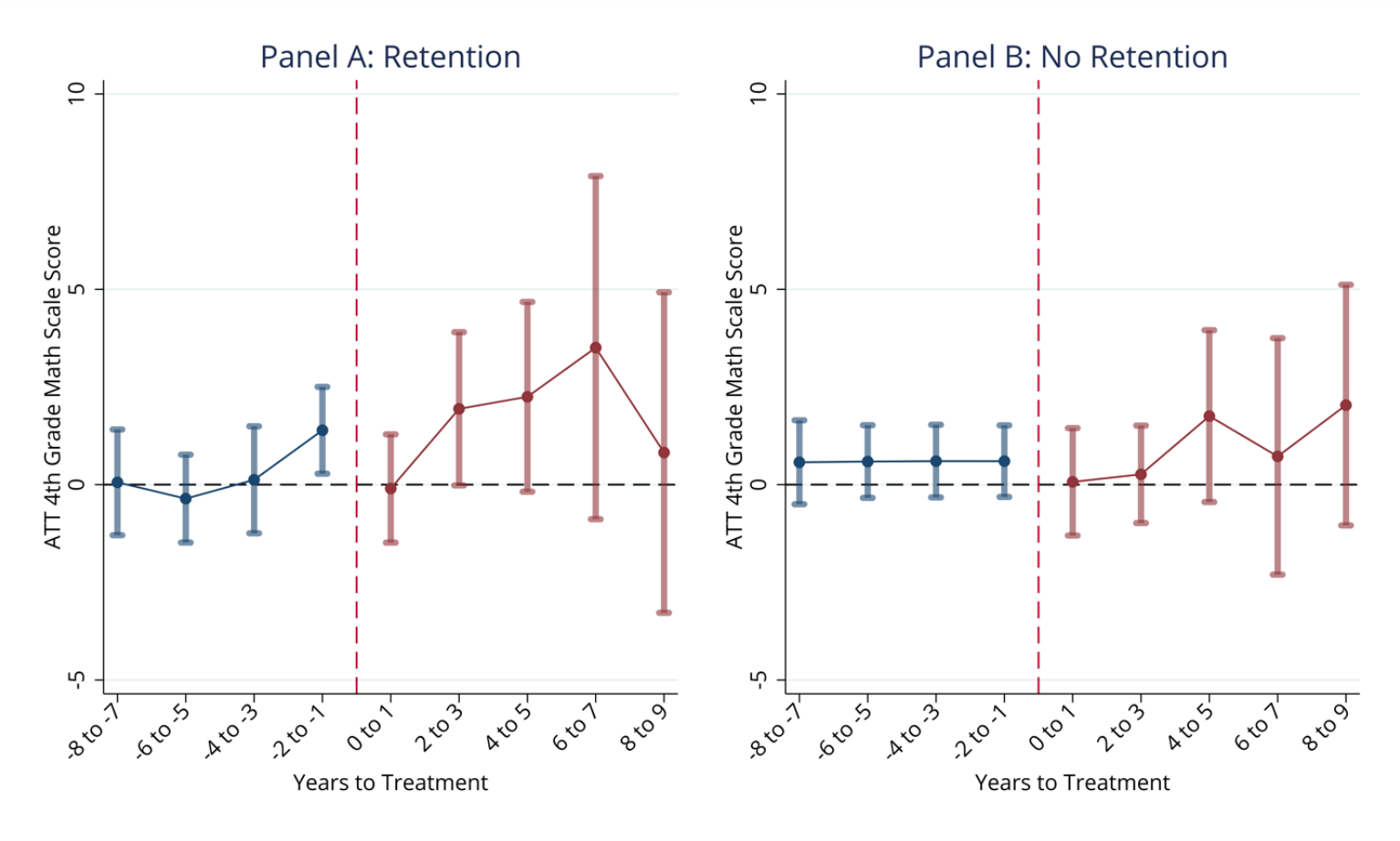
Notes: Data are from overall average NAEP fourth- and eighth-grade math scale scores, 2003-2019. We include only states that did not receive Race to the Top funding. Panel A and B examine fourth- and eighth-grade NAEP reading scores, respectively. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 15. Comprehensive Early Literacy Policies and Low-Stakes Math Scores



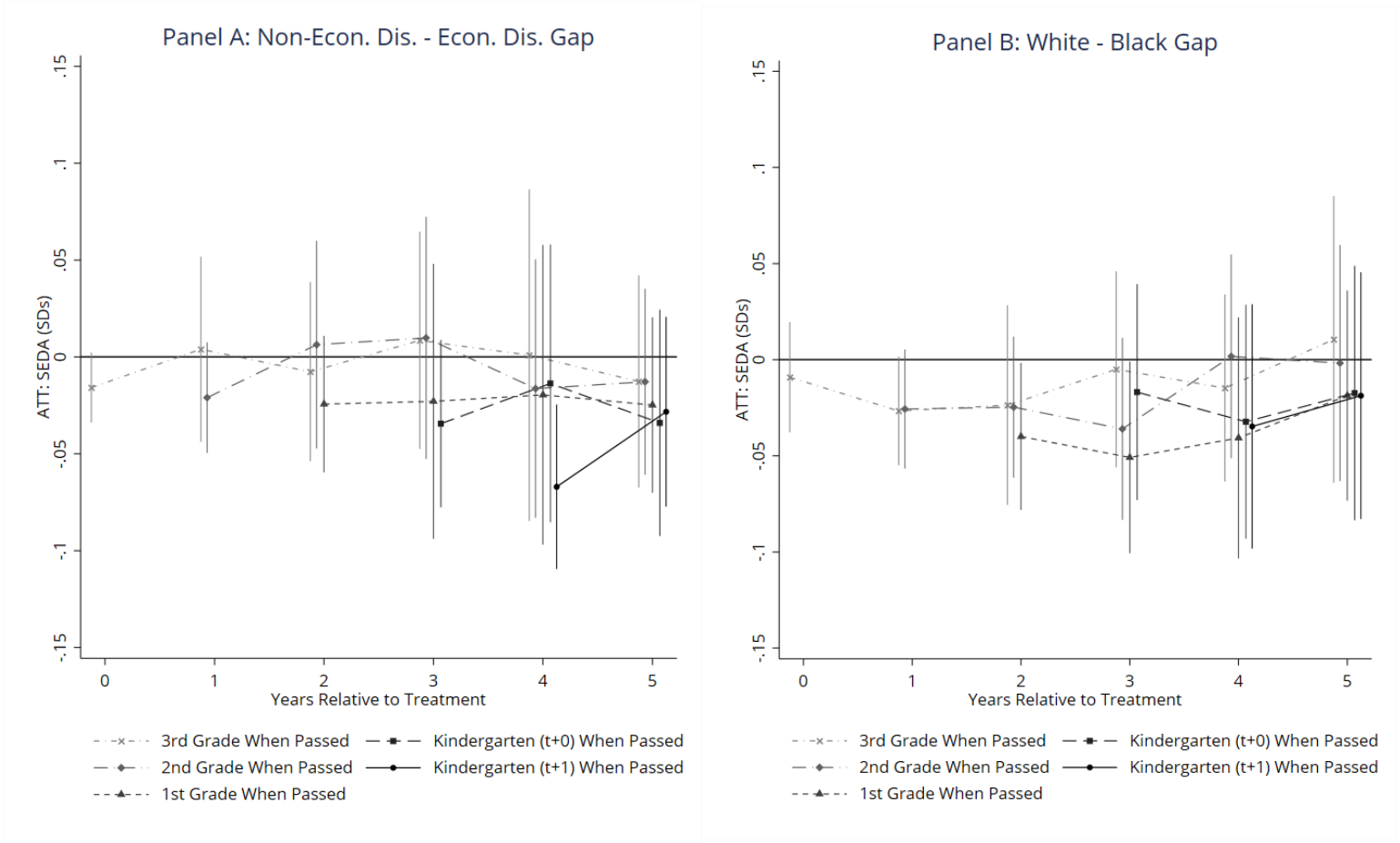
Note: Data are from overall average NAEP fourth-grade math scale scores, 2003-2019. We include only states that did not receive Race to the Top funding. Panel A includes states with comprehensive policies and never-treated states. Panel B includes states with non-comprehensive policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 16. Early Literacy Policies with Retention Requirements and Low-Stakes Math Scores



Note: Data are from overall average NAEP fourth-grade math scale scores, 2003-2019. We include only states that did not receive Race to the Top funding. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Figure 17. Early Literacy Policies and High-Stakes Reading Test Score Gaps



Note: Data are from average SEDA reading score gaps, 2009-2018. Panel A examines the non-economically disadvantaged-economically disadvantaged high-stakes reading gap. Panel B examines the White-Black high-stakes reading gap. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Table 1. Number of States Including Early Literacy Policy Components	
Policy Component	**Number of States**
Supports for Teachers & Policy	
Science of Reading Training	30
Literacy/Reading Coaches	23
Teacher Prep Program Alignment to SOR and/or SOR Assessment	38
Funding for Literacy Efforts	37
Assessment & Parent Notification	
Universal Screener Identify Students with Reading Deficiency (K-3)	39
Dyslexia Screener for At-Risk Students	22
Notify Parents of Students Identified with Reading Deficiency	32
Instruction & Intervention	
District Adoption of High-Quality Instructional Materials	24
Individual Reading Plan and/or Intervention for Students w/ a Reading Deficiency	38
Monitor Progress Students with Reading Deficiency (K-3)	36
Intervention During Summer/Before, During, and/or After School Hours	33
Summer Reading Camps/Innovative Summer Reading Programs	30
Parent Engagement At-Home Reading Strategies	29
Retention & Intensive Intervention	
Statewide: Initial Determinant Retention at 3rd Grade Based on State Assessment (Cut Score)	22
Multiple Options for Promotion	20
Good Cause Exemptions (GCEs) for Some Students	16

Note. Source: (ExcelinEd, 2021)

Table 2. SEDA and NAEP Sample Sizes								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	SEDA (2009-2018)						NAEP (2003-2019)	
	OVERALL		NON-ED-ED GAP		WHITE-BLACK GAP		OVERALL	
	States	State-Years	States	State-Years	States	State-Years	States	State-Years
	Panel A: Reading							
Grade 3	45	391	45	373	45	362	-	-
Grade 4	45	398	45	380	45	377	47	423
Grade 5	45	400	45	381	45	387	-	-
Grade 6	45	397	45	383	45	378	-	-
Grade 7	45	394	45	379	45	375	-	-
Grade 8	45	391	45	378	45	371	47	423
	Panel B: Math							
Grade 3	45	398	45	380	45	380	-	-
Grade 4	45	400	45	382	45	391	47	423
Grade 5	45	393	45	374	45	384	-	-
Grade 6	45	388	45	374	45	376	-	-
Grade 7	44	371	44	355	44	356	-	-
Grade 8	44	352	44	339	44	335	47	423

Note: These are the effective number of state-by-year observations used in our preferred event-study estimator (Callaway & Sant'Anna, 2021). These sample sizes exclude any states adopting early literacy policies before the analysis period. The SEDA data are yearly from 2009 to 2018. Some states are missing SEDA data in any given year. The NAEP data are every two years from 2003 to 2019. The NAEP data form a balanced panel.

Table 3. Outcome Data Summary Statistics

	SEDA						NAEP					
	Reading			Math			Reading			Math		
	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED
4th Grade												
Overall	0.012 (0.164)	0.021 (0.178)	-0.003 (0.139)	0.005 (0.193)	0.017 (0.201)	-0.016 (0.179)	220.935 (6.216)	221.175 (6.328)	220.179 (5.837)	240.343 (5.607)	240.433 (5.620)	240.059 (5.606)
White-Black Gap	0.717 (0.209)	0.770 (0.206)	0.628 (0.184)	0.772 (0.214)	0.824 (0.216)	0.684 (0.181)						
Non-Econ. Dis.-Econ. Dis. Gap	0.729 (0.122)	0.737 (0.121)	0.718 (0.122)	0.711 (0.134)	0.725 (0.134)	0.687 (0.131)						
8th Grade												
Overall	-0.003 (0.166)	0.010 (0.179)	-0.023 (0.142)	0.010 (0.202)	0.020 (0.208)	-0.007 (0.191)	264.932 (6.090)	264.857 (6.208)	265.170 (5.746)	282.660 (7.472)	282.422 (7.573)	283.410 (7.158)
White-Black Gap	0.684 (0.220)	0.744 (0.224)	0.587 (0.174)	0.752 (0.207)	0.815 (0.203)	0.647 (0.168)						
Non-Econ. Dis.-Econ. Dis. Gap	0.685 (0.108)	0.695 (0.110)	0.670 (0.104)	0.685 (0.114)	0.697 (0.117)	0.666 (0.106)						

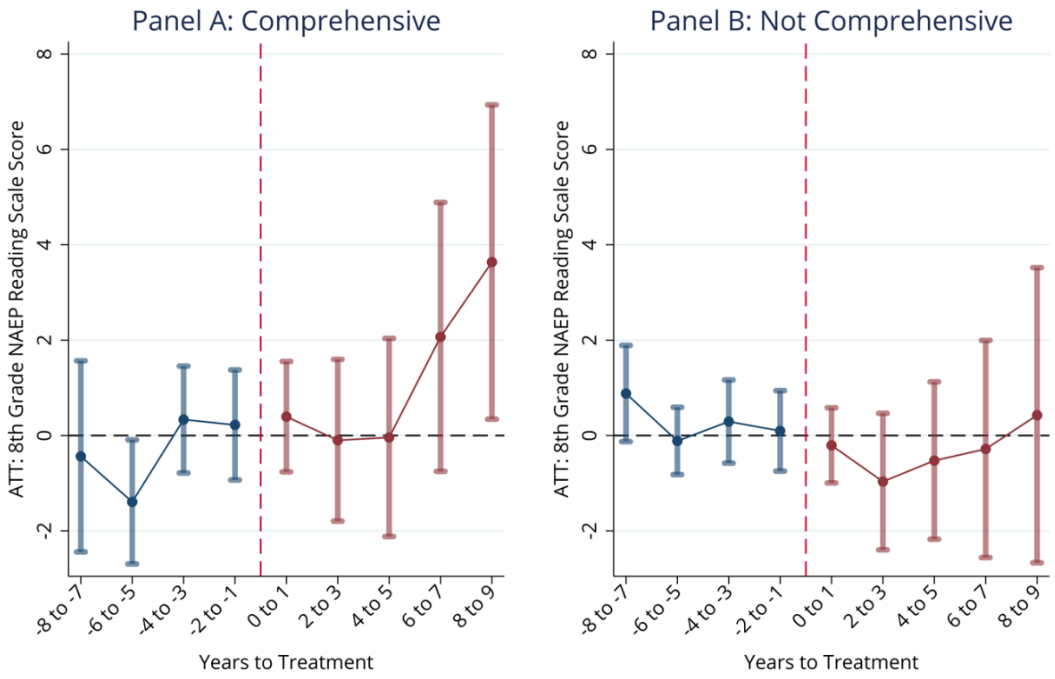
Note. These statistics are derived from NAEP and SEDA data from 2009 to 2018. The SEDA data are measured in standard deviations. The NAEP data are measured in scale score points. Standard deviations are in parentheses.

Table 4. Student Demographic Summary Statistics			
	Overall	Treated	Never Treated
% Urban	0.272	0.280	0.259
	(0.141)	(0.168)	(0.080)
% Suburb	0.305	0.317	0.287
	(0.190)	(0.199)	(0.175)
% Town	0.159	0.145	0.181
	(0.091)	(0.081)	(0.102)
% Rural	0.264	0.258	0.273
	(0.129)	(0.130)	(0.128)
Average Per-Grade Enrollment	69,637.98	56,856.07	89,926.52
	(84,759.47)	(34,484.73)	(126,626.82)
% Black	0.160	0.179	0.129
	(0.158)	(0.166)	(0.138)
% Asian	0.160	0.036	0.066
	(0.158)	(0.024)	(0.136)
% Hispanic	0.161	0.162	0.161
	(0.139)	(0.130)	(0.152)
% Native American	0.021	0.019	0.025
	(0.042)	(0.044)	(0.040)
% White	0.611	0.605	0.620
	(0.194)	(0.180)	(0.215)
% English Learner	0.064	0.065	0.063
	(0.043)	(0.034)	(0.054)
% Special Education	0.134	0.135	0.133
	(0.027)	(0.027)	(0.027)
% Economically Disadvantaged	0.502	0.497	0.509
	(0.112)	(0.108)	(0.118)
Unemployment Rate	0.079	0.082	0.075
	(0.027)	(0.028)	(0.027)
Poverty Rate	0.143	0.142	0.143
	(0.033)	(0.034)	(0.031)
BA+ Rate	0.295	0.301	0.286
	(0.061)	(0.069)	(0.042)
Single-Mother Household Rate	0.188	0.192	0.182
	(0.043)	(0.047)	(0.037)

Note: Data are from the SEDA. State-level averages from 2009 to 2018. Standard deviations are in parentheses.

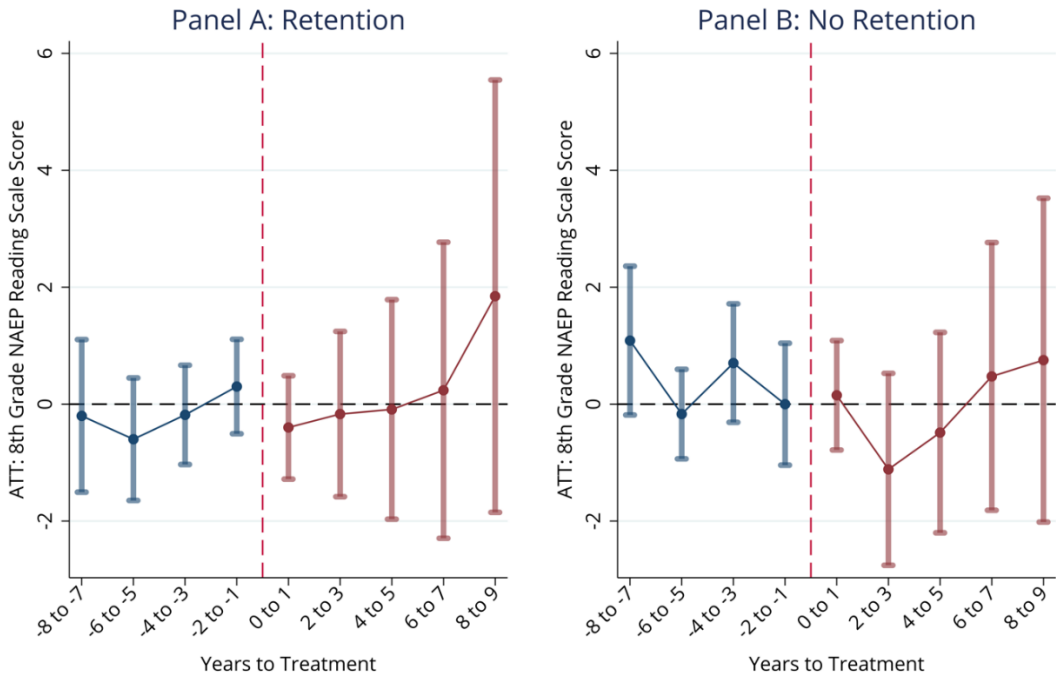
APPENDIX

Appendix Figure A1. Comprehensive Early Literacy Policies and 8th Low-Stakes Reading Scores



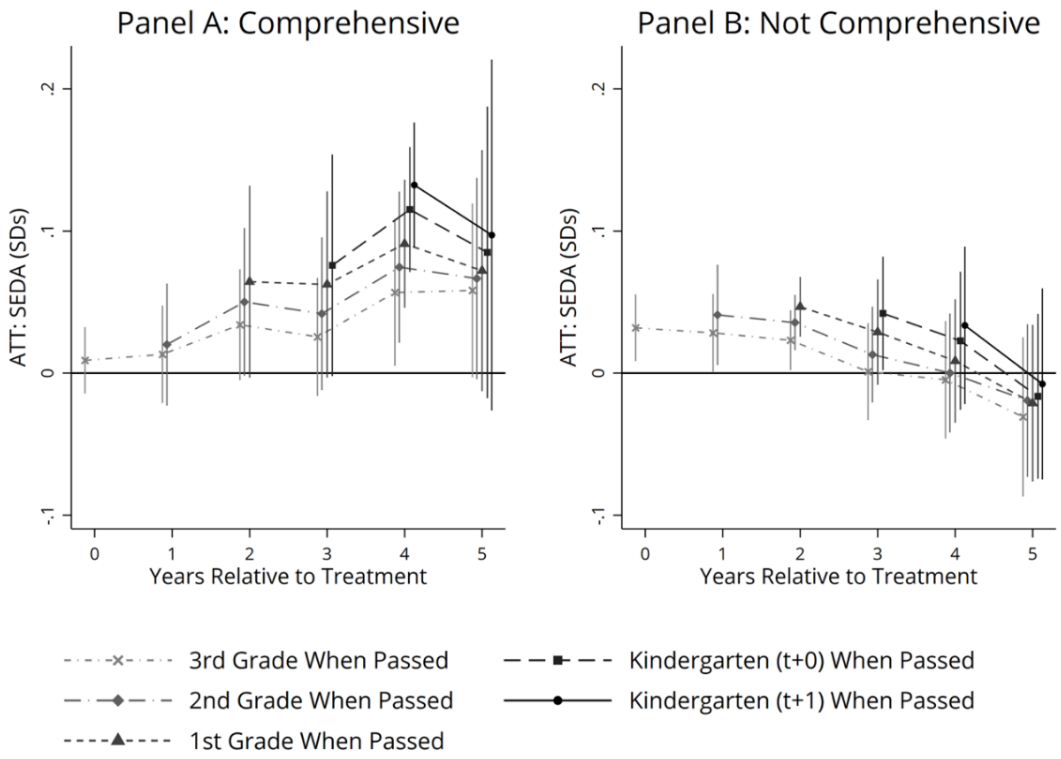
Note: Data are from overall average NAEP eighth-grade reading scale scores, 2003-2019. Panel A includes states with comprehensive policies and never-treated states. Panel B includes states with non-comprehensive policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A2. Early Literacy Policies with Retention Requirements and 8th Low-Stakes Reading Scores



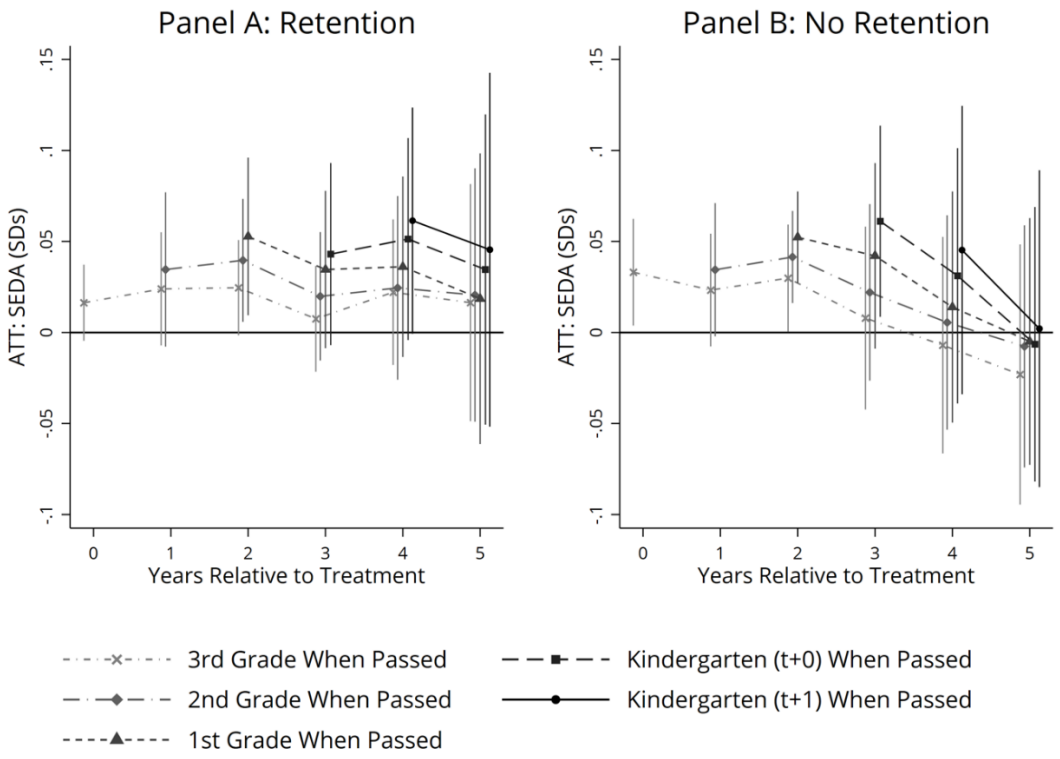
Note: Data are from overall average NAEP eighth-grade reading scale scores, 2003-2019. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A3. No Race to the Top Funding, Comprehensive Early Literacy Policies, and High-Stakes Reading Scores



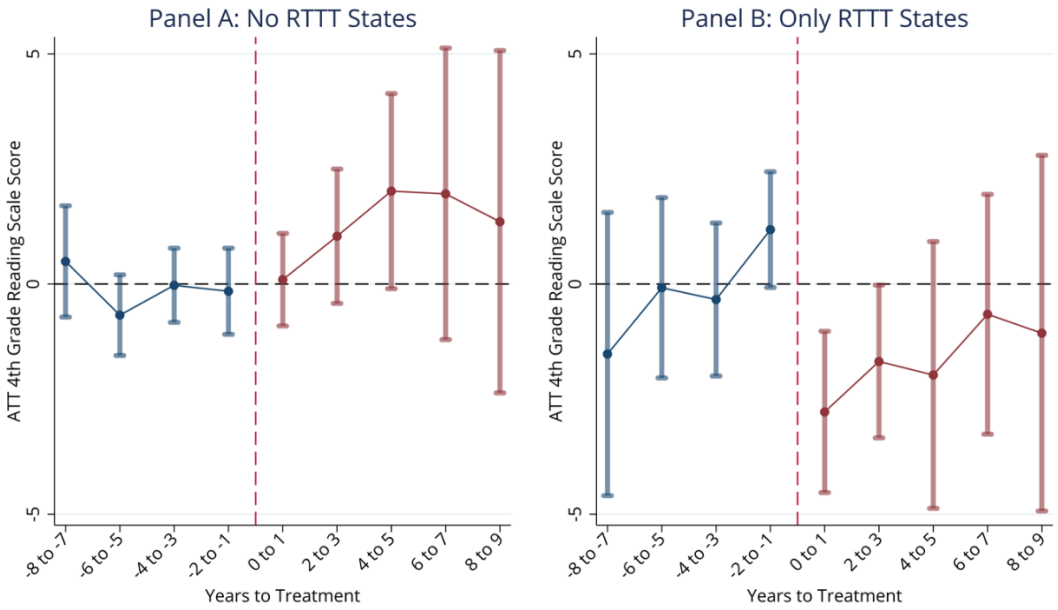
Note: Data are from overall average SEDA reading scores, 2009-2018. We exclude states that received Race to the Top funding. Panel A includes states with comprehensive early literacy policies and never-treated states. Panel B includes states without comprehensive early literacy policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A4. No Race to the Top Funding, Early Literacy Policies with Retention Requirements, and High-Stakes Reading Scores



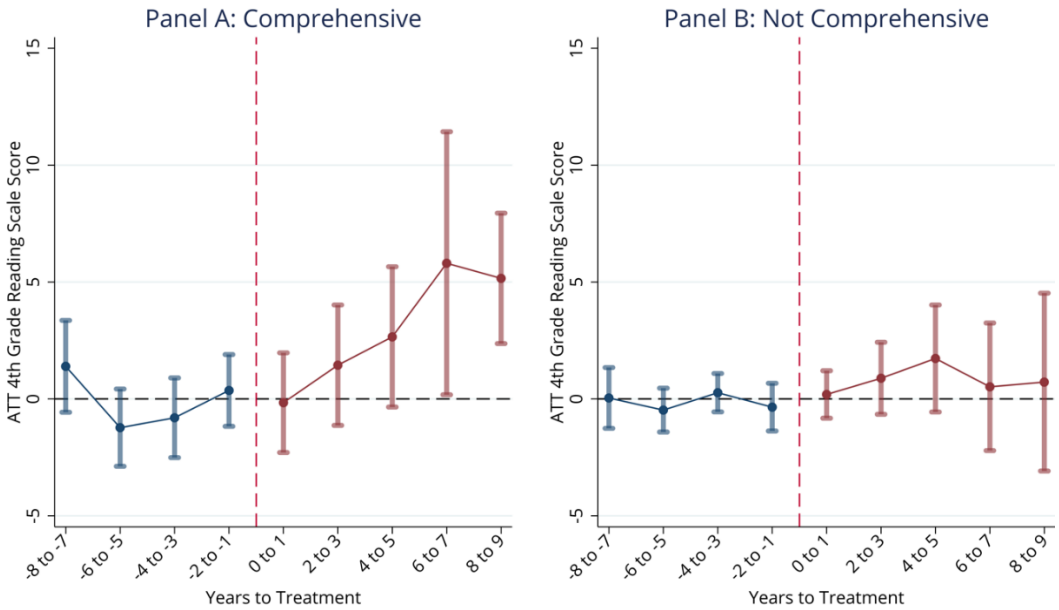
Note: Data are from overall average SEDA reading scores, 2009-2018. We exclude states that received Race to the Top funding. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Table A5. Race to the Top Funding, Any Early Literacy Policy, and 4th Low-Stakes Reading Scores



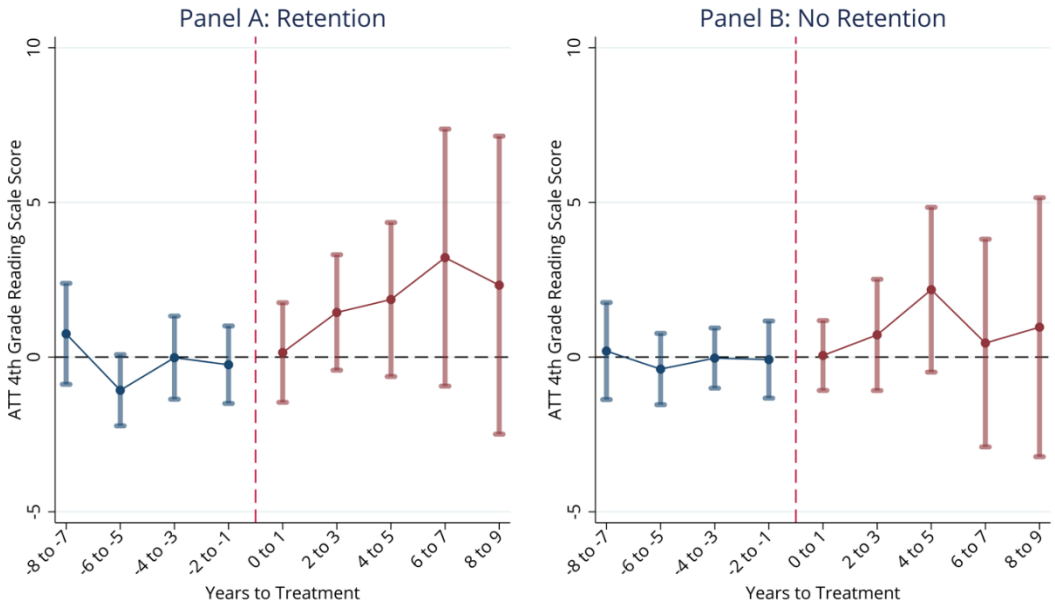
Note: Data are from overall average NAEP fourth-grade reading scale scores, 2003-2019. Panel A includes states that did not receive Race to the Top funding. Panel B includes states that received Race to the Top funding. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A6. No Race to the Top Funding, Comprehensive Early Literacy Policies, and 4th Low-Stakes Reading Scores



Note: Data are from overall average NAEP fourth-grade reading scale scores, 2003-2019. We exclude states that received Race to the Top funding. Panel A includes states with comprehensive policies and never-treated states. Panel B includes states with non-comprehensive policies and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level

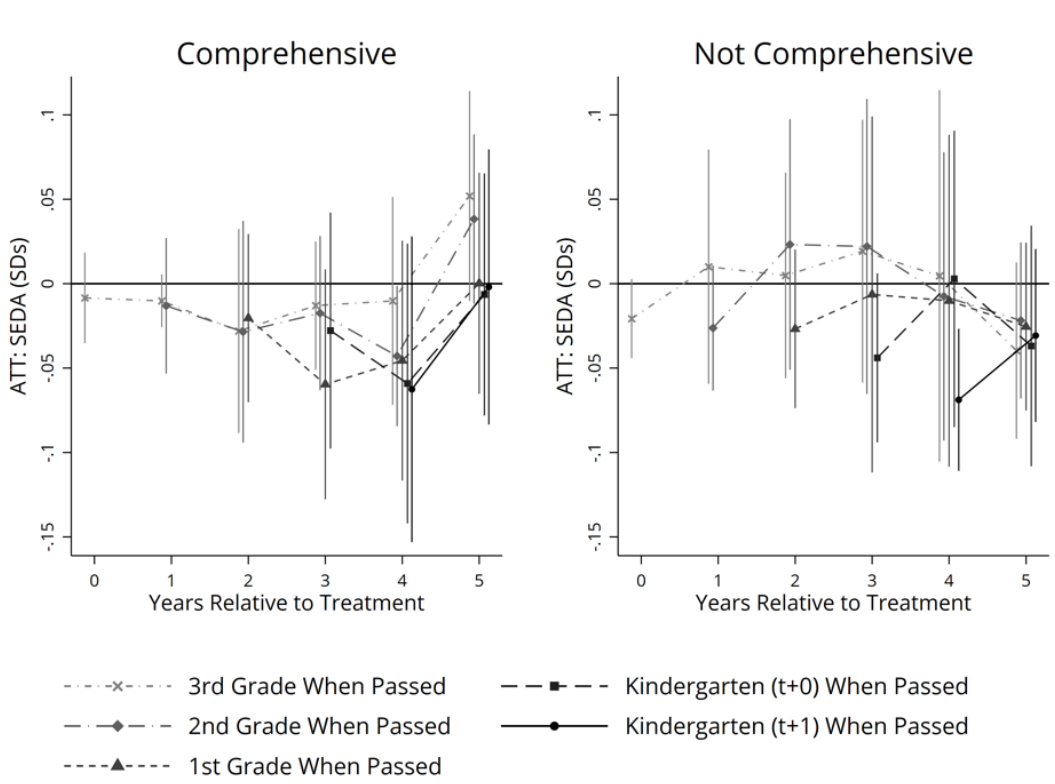
Appendix Figure A7. No Race to the Top Funding, Early Literacy Policies with Retention Requirements and 4th Low-Stakes Reading Scores



Note: Data are from overall average NAEP fourth-grade reading scale scores, 2003-2019. We exclude states that received Race to the Top funding. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A8. Comprehensive Early Literacy Policies and High-Stakes Achievement Gaps

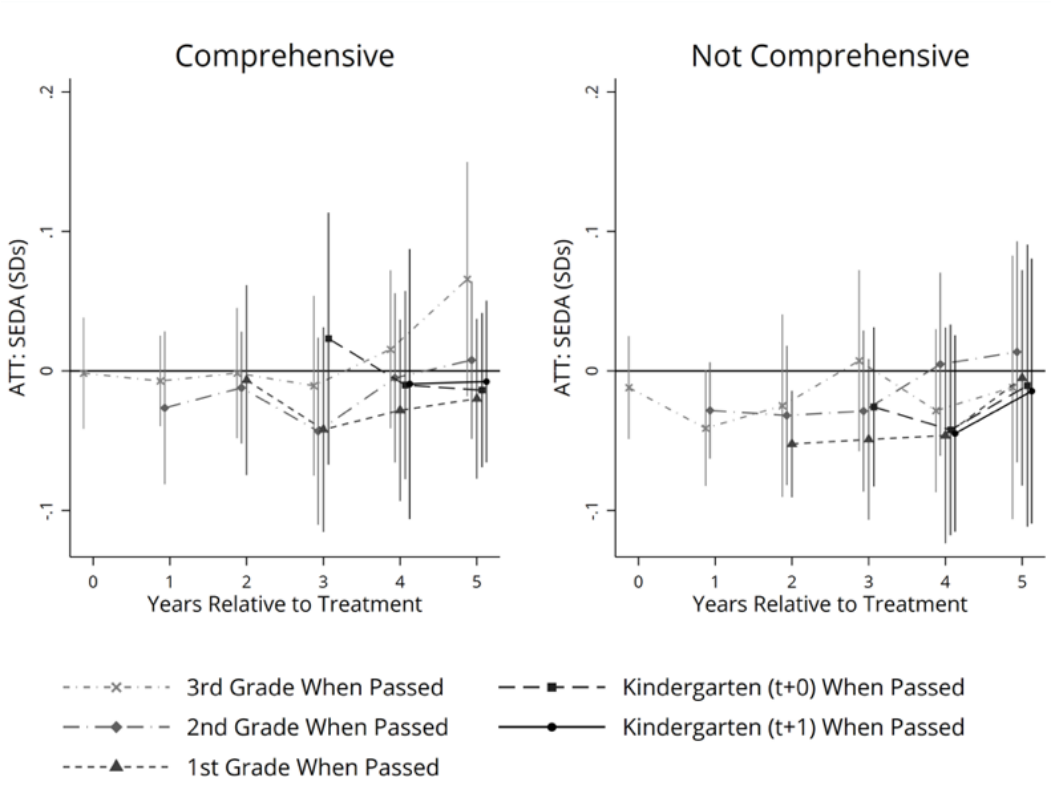
Non-Economically Disadvantaged-Economically Disadvantaged Gap



Note: Data are from SEDA non-economically disadvantaged-economically disadvantaged reading score gaps, 2009-2018. Panel A includes states with comprehensive early literacy policies and never-treated states. Panel B includes states without comprehensive early literacy policies and never-treated states. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A9. Comprehensive Early Literacy Policies and High-Stakes Achievement Gaps

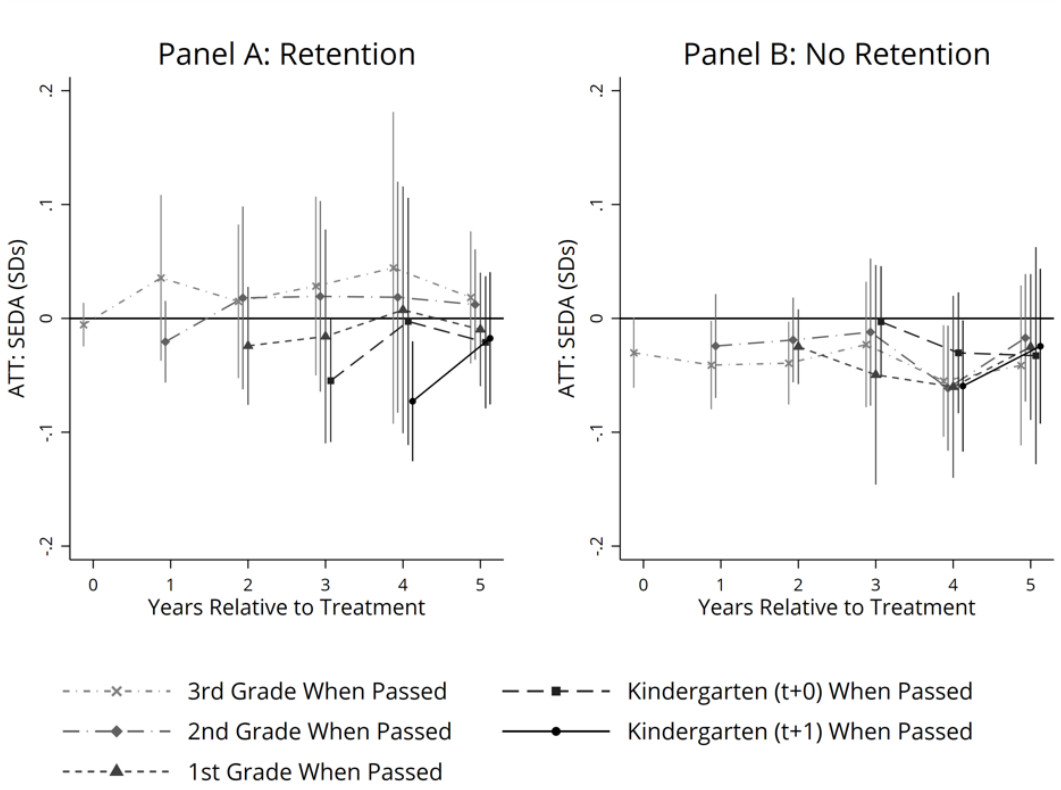
White-Black Gap



Note: Data are from SEDA White-Black reading score gaps, 2009-2018. Panel A includes states with comprehensive early literacy policies and never-treated states. Panel B includes states without comprehensive early literacy policies and never-treated states. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A10. Early Literacy Policies with Retention Requirements and High-Stakes Achievement Gaps

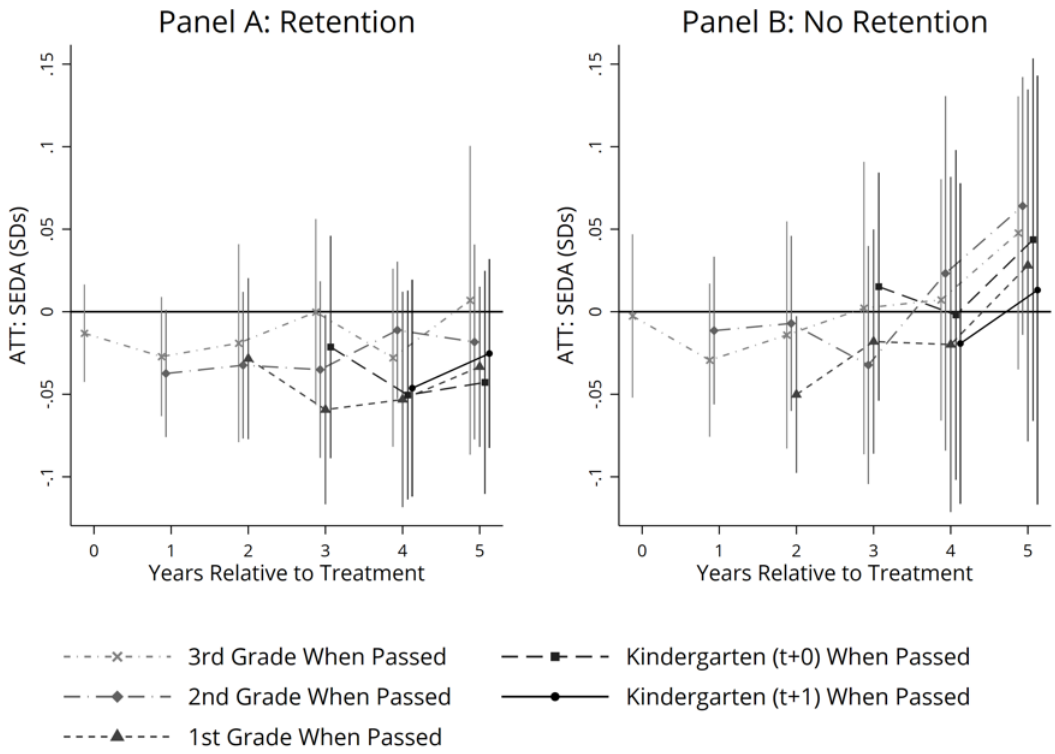
Non-Economically Disadvantaged-Economically Disadvantaged Gap



Note: Data are from SEDA non-economically disadvantaged-economically disadvantaged reading score gaps, 2009-2018. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

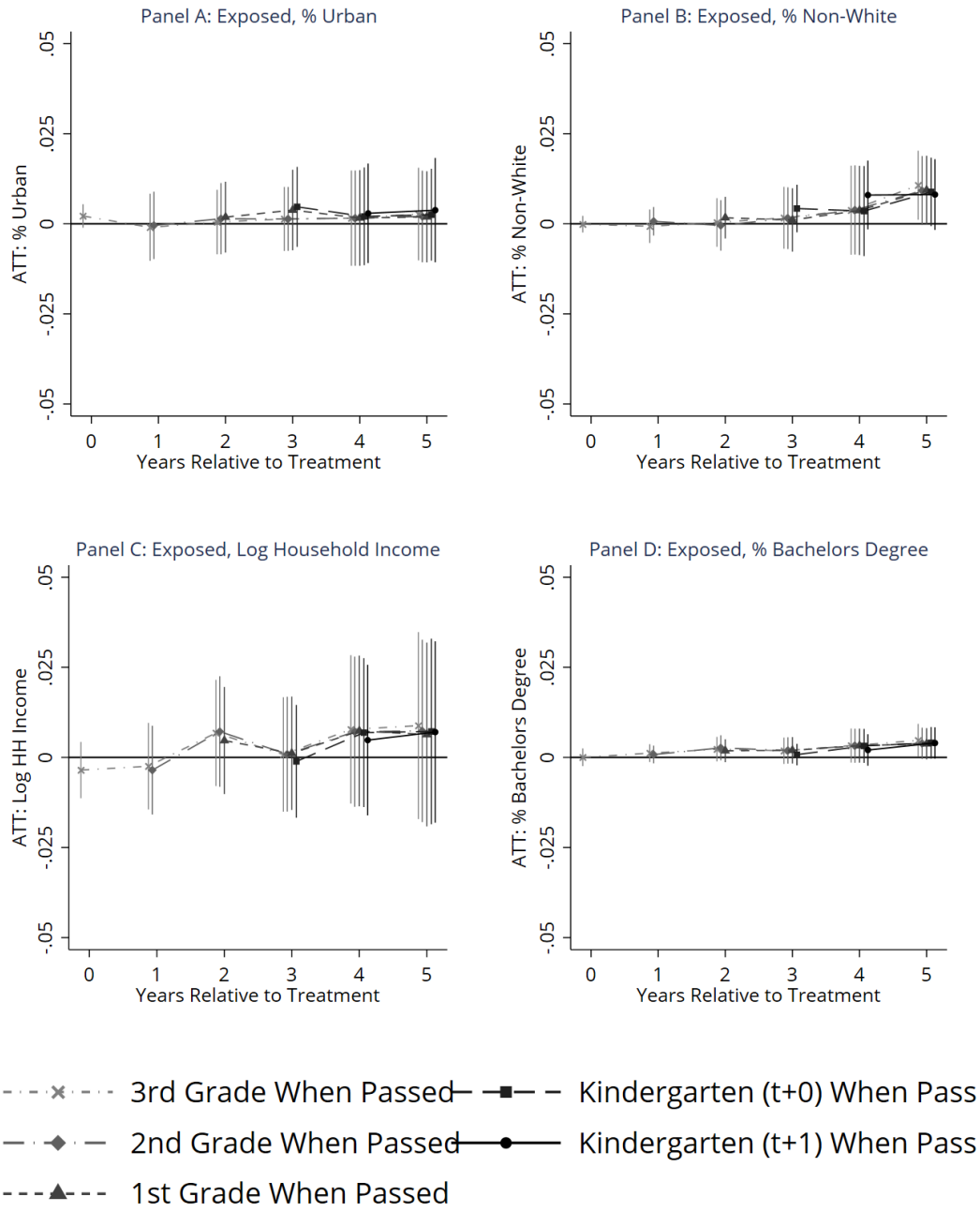
Appendix Figure A11. Early Literacy Policies with Retention Requirements and High-Stakes Achievement Gaps

White-Black Gap



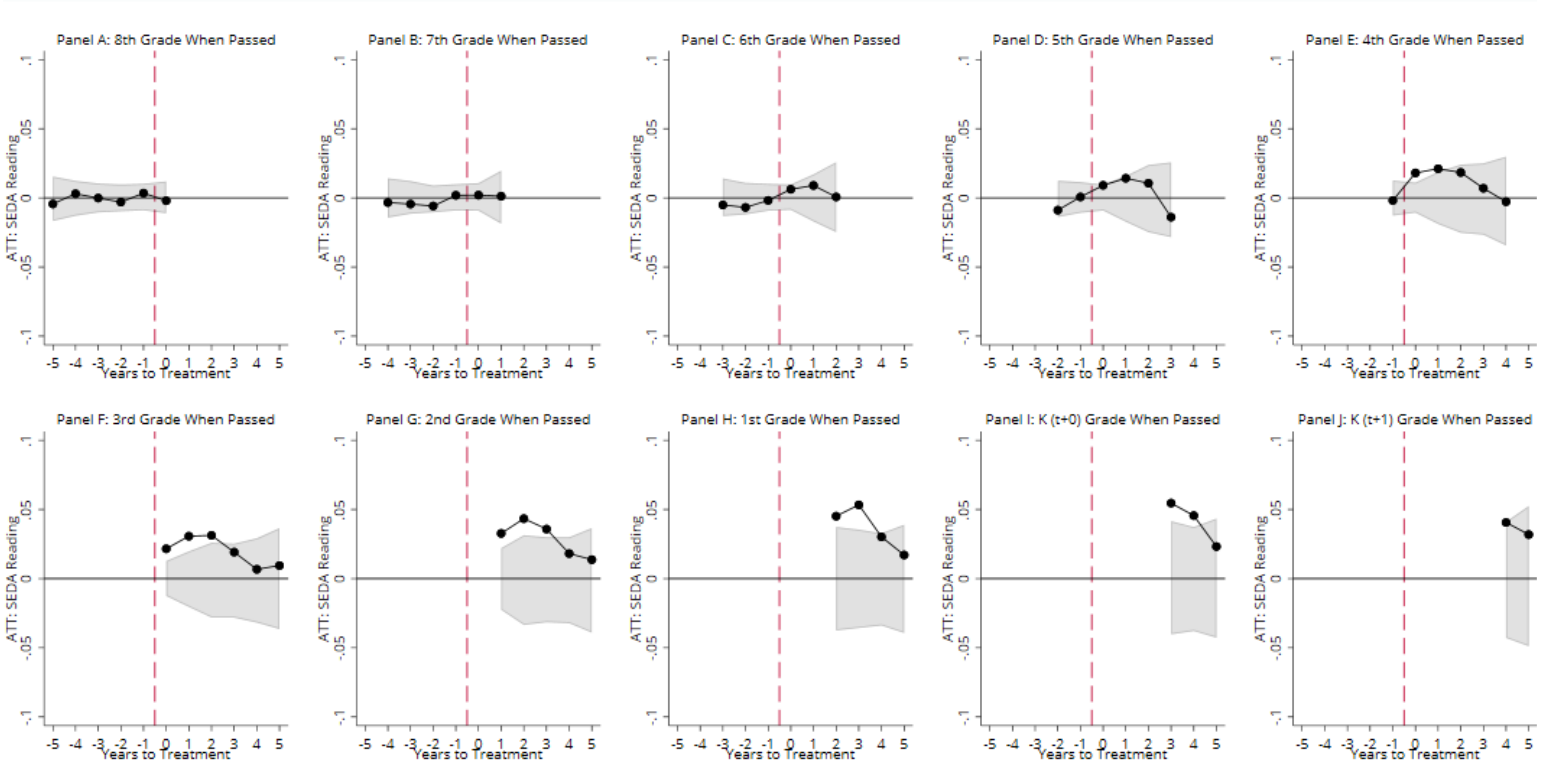
Note: Data are from SEDA White-Black reading score gaps, 2009-2018. Panel A includes states with retention requirements and never-treated states. Panel B includes states without retention requirements and never-treated states. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A12. Falsification Tests



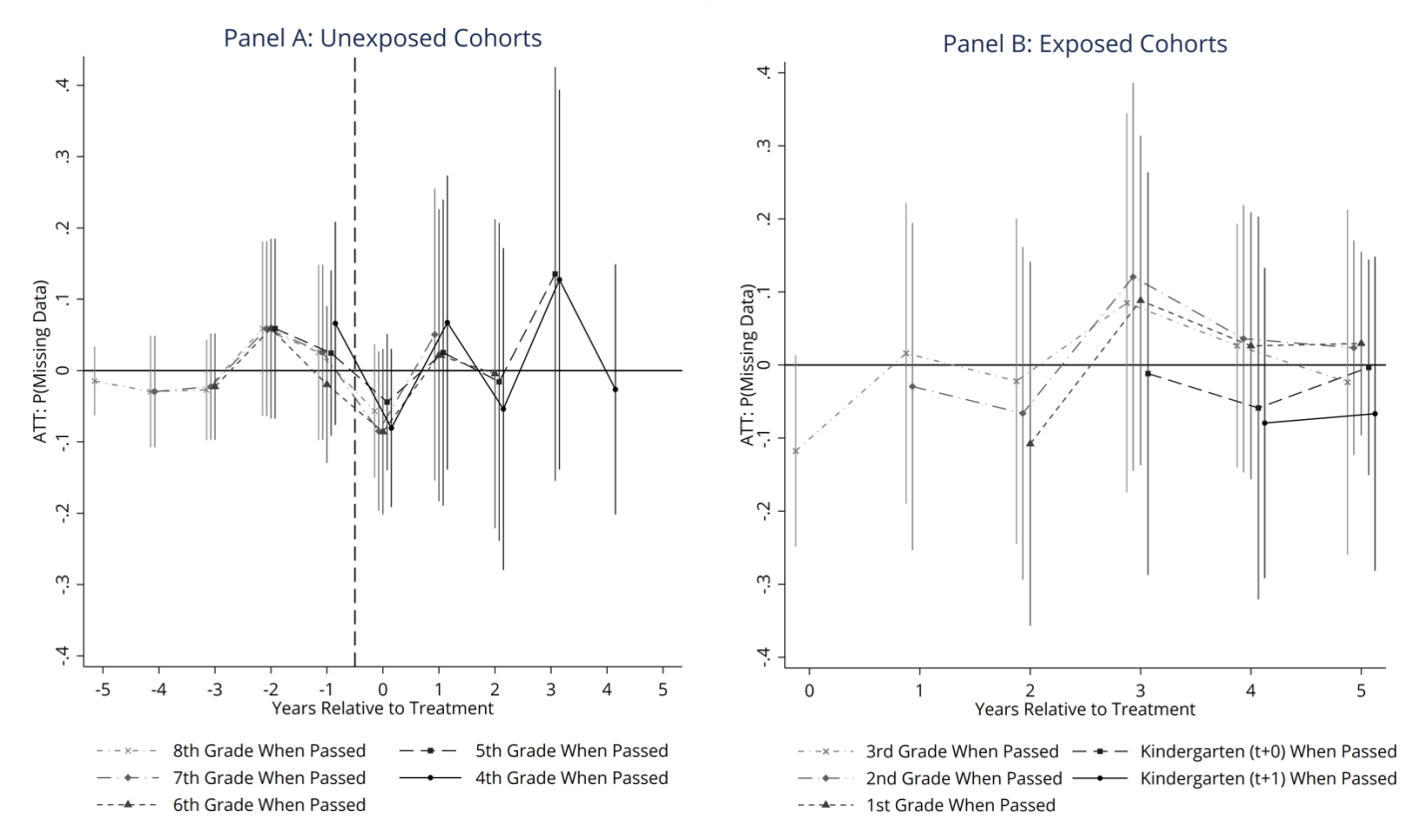
Note: Data are from the SEDA, 2009-2018. In Panel A, the outcome is the percent urban in a state-year. In Panel B, the outcome is percent non-white in a state-year. In Panel C, the outcome is log median household income in a state-year. In Panel D, the outcome is percent with a bachelor's degree or higher in a state-year. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. Relative years binned into two-year bins due to the biennial nature of NAEP administration. 95% confidence intervals from standard errors clustered at the state level.:

Appendix Figure A13. Placebo Test – Random Treatment Assignment (High-Stakes SEDA Reading)



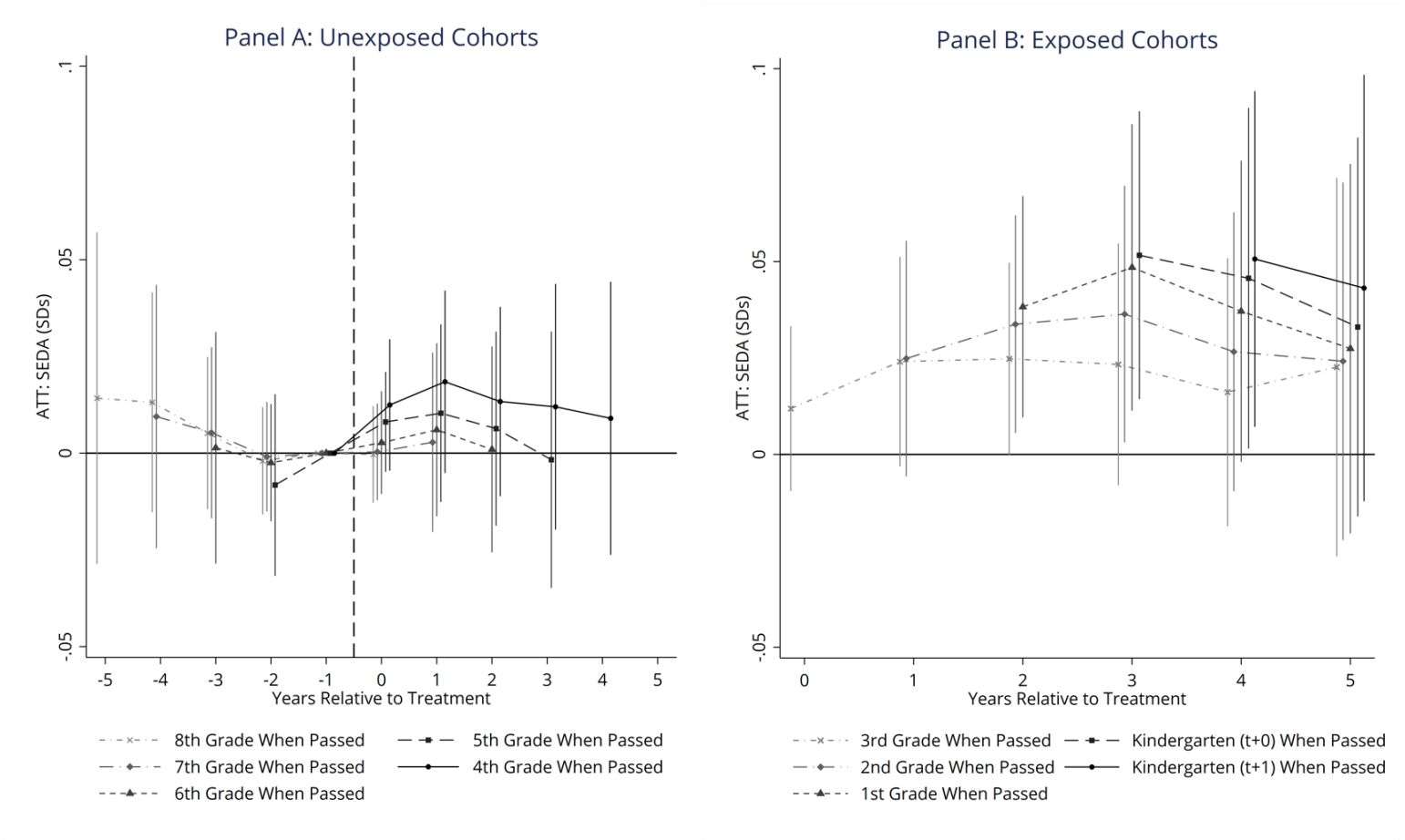
Note: Data are from the third- through eighth-grade overall average SEDA reading scores, 2009-2018. The shaded region represents the empirical 95% confidence interval constructed by randomly assigning treatment status to states 500 times (keeping the time distribution of adoption constant). Line represents the actual event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator shown in Figure 3, Panel A.

Appendix Figure A14. Missing Data Analysis



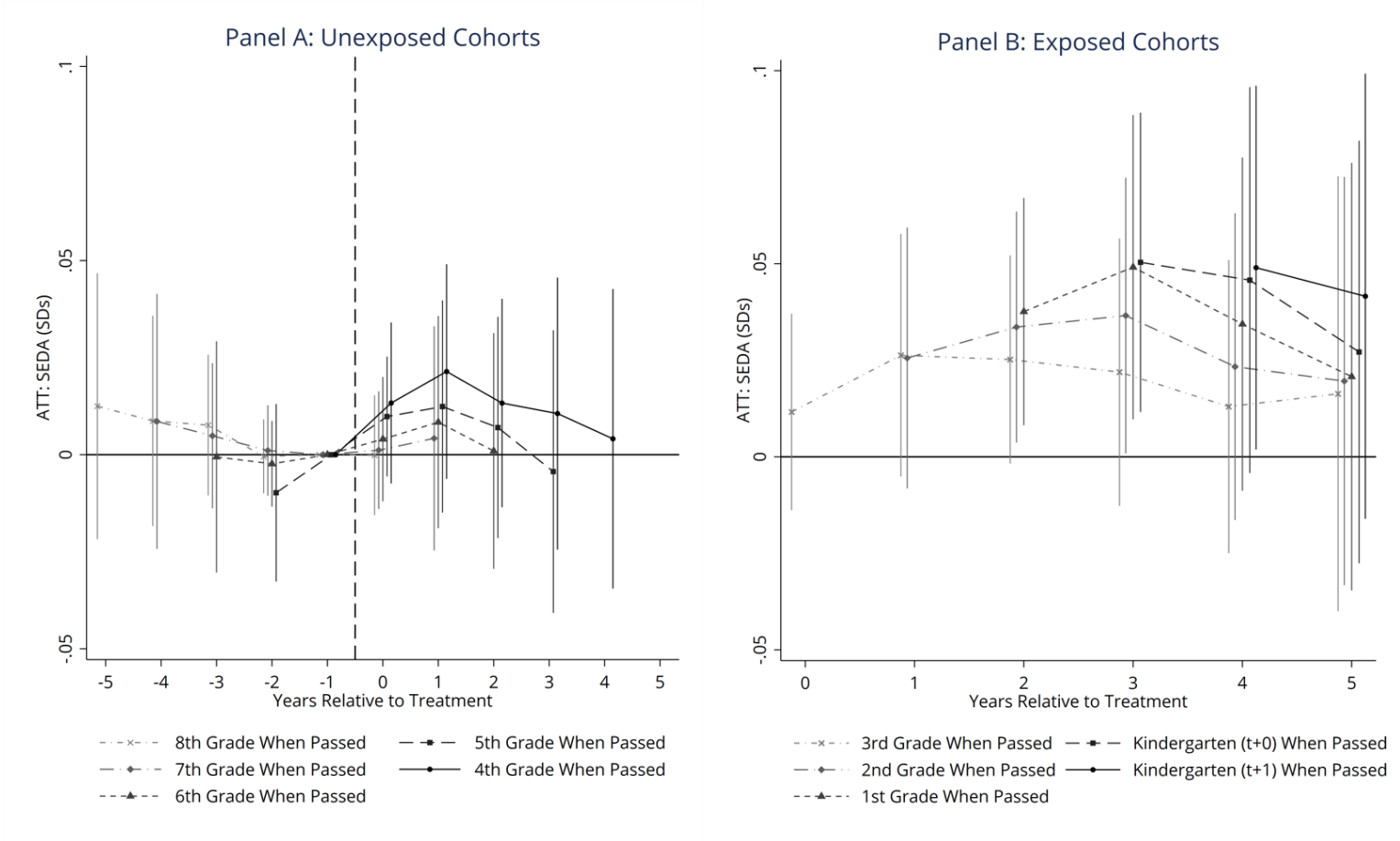
Note: Data are derived from overall average SEDA reading scores, 2009-2018. The outcome is an indicator for missing data in a given grade-by-state-by-year. Event study coefficient estimates from the Callaway & Sant'Anna (2021) estimator. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A15. Two-Way Fixed Effects Event Study (OLS)



Note: Data are from overall average SEDA reading scores, 2009-2018. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from a two-way fixed effects event study model estimated using OLS. 95% confidence intervals from standard errors clustered at the state level.

Appendix Figure A16. Interaction-Weighted Event-Study Estimator (Sun & Abraham, 2021)



Note: Data are from overall average SEDA reading scores, 2009-2018. The sample sizes range from 391 to 400 state-year observations. Detailed sample sizes can be found in Table 2. Event study coefficient estimates from the interaction-weighted event-study estimator (Sun & Abraham, 2021). 95% confidence intervals from standard errors clustered at the state level.

Appendix Table A1. Outcome Data Summary Statistics												
	SEDA						NAEP					
	Reading			Math			Reading			Math		
	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED
3rd Grade												
Overall	0.023 (0.166)	0.033 (0.180)	0.007 (0.141)	0.008 (0.196)	0.023 (0.203)	-0.017 (0.182)						
White-Black Gap	0.693 (0.185)	0.740 (0.177)	0.617 (0.174)	0.765 (0.203)	0.809 (0.206)	0.692 (0.175)						
Non-Econ. Dis.-Econ. Dis. Gap	0.719 (0.119)	0.723 (0.119)	0.712 (0.120)	0.707 (0.129)	0.717 (0.129)	0.691 (0.128)						
4th Grade												
Overall	0.012 (0.164)	0.021 (0.178)	-0.003 (0.139)	0.005 (0.193)	0.017 (0.201)	-0.016 (0.179)	220.94 (6.216)	221.18 (6.328)	220.18 (5.837)	240.34 (5.607)	240.43 (5.620)	240.06 (5.606)
White-Black Gap	0.717 (0.209)	0.770 (0.206)	0.628 (0.184)	0.772 (0.214)	0.824 (0.216)	0.684 (0.181)						
Non-Econ. Dis.-Econ. Dis. Gap	0.729 (0.122)	0.737 (0.121)	0.718 (0.122)	0.711 (0.134)	0.725 (0.134)	0.687 (0.131)						
5th Grade												
Overall	0.002 (0.163)	0.012 (0.176)	-0.014 (0.139)	0.000 (0.193)	0.011 (0.201)	-0.017 (0.180)						
White-Black Gap	0.706 (0.209)	0.765 (0.204)	0.612 (0.182)	0.755 (0.207)	0.811 (0.205)	0.665 (0.178)						
Non-Econ. Dis.-Econ. Dis. Gap	0.725 (0.119)	0.736 (0.118)	0.708 (0.119)	0.702 (0.121)	0.716 (0.121)	0.681 (0.119)						

Appendix Table A1. Outcome Data Summary Statistics												
	SEDA						NAEP					
	Reading			Math			Reading			Math		
	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED	OVERALL	TREATED	NEVER TREATED
6th Grade												
Overall	-0.008 (0.163)	0.002 (0.176)	-0.023 (0.140)	-0.003 (0.197)	0.007 (0.206)	-0.017 (0.183)						
White-Black Gap	0.709 (0.213)	0.767 (0.207)	0.616 (0.188)	0.771 (0.217)	0.832 (0.217)	0.674 (0.179)						
Non-Econ. Dis.-Econ. Dis. Gap	0.725 (0.106)	0.731 (0.107)	0.717 (0.105)	0.713 (0.114)	0.726 (0.115)	0.694 (0.111)						
7th Grade												
Overall	-0.006 (0.165)	0.003 (0.178)	-0.021 (0.143)	0.001 (0.203)	0.007 (0.211)	-0.010 (0.189)						
White-Black Gap	0.704 (0.232)	0.767 (0.238)	0.601 (0.180)	0.770 (0.228)	0.838 (0.227)	0.649 (0.173)						
Non-Econ. Dis.-Econ. Dis. Gap	0.713 (0.105)	0.720 (0.106)	0.702 (0.103)	0.708 (0.115)	0.719 (0.118)	0.690 (0.108)						
8th Grade												
Overall	-0.003 (0.166)	0.010 (0.179)	-0.023 (0.142)	0.010 (0.202)	0.020 (0.208)	-0.007 (0.191)	264.93 (6.090)	264.86 (6.208)	265.17 (5.746)	282.66 (7.472)	282.42 (7.573)	283.41 (7.158)
White-Black Gap	0.684 (0.220)	0.744 (0.224)	0.587 (0.174)	0.752 (0.207)	0.815 (0.203)	0.647 (0.168)						
Non-Econ. Dis.-Econ. Dis. Gap	0.685 (0.108)	0.695 (0.110)	0.670 (0.104)	0.685 (0.114)	0.697 (0.117)	0.666 (0.106)						

Note. These statistics are derived from NAEP and SEDA data from 2009 to 2018. The SEDA data are measured in standard deviations. The NAEP data are measured in scale score points. Standard deviations are in parentheses.