Do State Dyslexia Laws Improve Reading Achievement? Evidence from Staggered Adoption with Corrected Treatment Timing

APEP Autonomous Research*

January 25, 2026

Abstract

Since 1995, 27 U.S. states have adopted dyslexia-related legislation ranging from awareness provisions to comprehensive dyslexia legislation (26 since 2010, plus Texas in 1995). This paper evaluates the causal effect of these policies on fourth-grade reading achievement using a staggered difference-in-differences design with NAEP data from 2003–2022. A critical methodological contribution is correcting treatment timing: because NAEP is administered January–March, laws effective mid-year cannot affect that year's assessment. Employing the Callaway-Sant'Anna (2021) estimator with corrected timing and 1,000 bootstrap iterations, I find a precisely estimated null average effect: ATT = 1.02 NAEP points (SE = 1.16). However, separating states that bundled dyslexia laws with comprehensive literacy reforms from "dyslexia-only" states reveals important heterogeneity. Among bundled reform states with evaluable post-treatment NAEP data (Mississippi, Florida, Tennessee), positive effects emerge, while dyslexia-only mandates show null effects. Alabama also adopted a bundled reform (2022) but has no post-treatment NAEP observation in the sample period. These findings suggest that dyslexia legislation alone—without accompanying curriculum reform, teacher training, and intervention requirements—do not improve aggregate reading outcomes. The policy implication is clear: effective early literacy policy requires comprehensive reform bundles, not piecemeal mandates.

JEL Codes: I21, I28, H75

Keywords: dyslexia legislation, education policy, reading achievement, difference-in-differences, NAEP, treatment timing

[*] Autonomous Policy Evaluation Project. This paper was autonomously generated using Claude Code. Revision of apep_0069. Project Repository: https://github.com/SocialCatalystLab/auto-policy-evals

1. Introduction

Reading proficiency in elementary school is among the strongest predictors of long-term educational and economic success. Students who fail to read proficiently by third grade are four times more likely to drop out of high school (Hernandez, 2011), and early reading difficulties often persist into adulthood, affecting employment prospects and lifetime earnings (Kutner et al., 2007). Dyslexia—a neurobiological learning disability affecting 5–10% of the population—is a leading cause of reading difficulties (Shaywitz, 2003). Yet dyslexia often goes undiagnosed until students have fallen significantly behind, when intervention becomes more difficult and less effective.

In response to these concerns, state legislatures have adopted dyslexia-related legislation at an accelerating pace. Between 2010 and 2022, 26 states adopted laws ranging from minimal awareness provisions to comprehensive mandates requiring universal screening, evidence-based intervention, teacher training, and dedicated funding. The theory of change is straightforward: legislation that promotes screening identifies struggling readers earlier, enabling targeted intervention before skills deficits compound. This staggered adoption across states creates natural variation that can be exploited for causal identification.

This paper makes three contributions to the literature on early literacy policy. First, I correct a critical measurement issue: the timing of treatment exposure relative to outcome measurement. NAEP assessments are administered in January–March of each year, but most dyslexia laws become effective on July 1 or later. A law "effective in 2019" cannot possibly affect the 2019 NAEP assessment because implementation began after testing concluded. The first NAEP that could reflect the policy's effects is the *next* assessment wave. Failing to account for this timing mismatch mechanically attenuates treatment effects toward zero because "treated" observations are actually pre-treatment. Prior research on state education mandates has generally not addressed this issue.

Second, I distinguish between two types of policy intervention: (1) "dyslexia-only" laws that focus on dyslexia-specific provisions (awareness, screening, intervention requirements) without accompanying system-wide literacy reforms, and (2) "bundled reform" packages that combine dyslexia provisions with comprehensive Science of Reading reforms, including structured literacy curricula, phonics-based teacher training, and in some cases third-grade retention policies. Mississippi's 2013 Literacy-Based Promotion Act is the paradigmatic example of a bundled reform, combining dyslexia provisions with curriculum overhaul, intensive teacher training, and grade retention. Florida, Tennessee, and Alabama have adopted similar comprehensive packages. Pooling these states with those that adopted dyslexia-focused legislation obscures the distinction between policy mechanisms.

Third, I implement modern heterogeneity-robust difference-in-differences methods with careful attention to inference. Using the Callaway-Sant'Anna (2021) estimator with 1,000 bootstrap iterations, I estimate group-time average treatment effects that avoid the well-documented biases of two-way fixed effects in staggered settings (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021). I report formal pre-trend tests and binned event studies to address sparse cells at extreme event times.

The main findings can be summarized as follows. The average treatment effect on the treated across all states with dyslexia laws is approximately zero: ATT = 1.02 NAEP scale points (SE = 1.16), with a 95% confidence interval [-1.26, 3.30] that cannot distinguish the effect from zero. This null result is robust across specifications and passes placebo tests using unrelated outcomes (Grade 4 math, Grade 8 reading).

However, the aggregate null masks important heterogeneity. When I separately estimate effects for bundled reform states with evaluable post-treatment NAEP (Mississippi, Florida, Tennessee) versus dyslexia-only states, a pattern emerges: bundled reform states show positive effects of approximately 3.2 NAEP scale points (SE = 2.1), while dyslexia-only states show null effects (ATT = 0.68, SE = 0.86). Alabama also adopted a bundled reform in 2022 but has no post-treatment NAEP in the sample. The bundled reform estimate is imprecise due to the small number of evaluable states (three), but the pattern suggests that comprehensive reform bundles may be more effective than dyslexia legislation alone.

These findings have immediate policy implications. State legislators considering dyslexia legislation should recognize that dyslexia-focused laws alone—even those with screening requirements—are unlikely to improve aggregate reading outcomes. Effective policy requires a comprehensive approach: mandatory screening to identify struggling readers, evidence-based intervention to address their needs, teacher training to ensure instructional quality, and adequate funding to support implementation. Piecemeal legislation without accompanying resources and systemic reforms may create compliance burdens without corresponding benefits.

The remainder of the paper proceeds as follows. Section 2 provides detailed institutional background on dyslexia legislation, including a treatment classification table documenting each state's law components and effective dates. Section 3 reviews related literature. Section 4 describes data sources and the critical treatment timing correction. Section 5 presents the empirical strategy. Section 6 reports results, including the bundled/dyslexia-only decomposition and robustness checks. Section 7 discusses implications and limitations. Section 8 concludes.

2. Institutional Background

2.1 Dyslexia and the Case for Early Intervention

Dyslexia is a specific learning disability that is neurobiological in origin, affecting approximately 5–10% of the population (Shaywitz, 2003). It is characterized by difficulties with accurate and fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language and are often unexpected given the individual's cognitive abilities and educational opportunities.

The scientific case for early intervention is compelling. Brain imaging studies demonstrate that the neural circuitry underlying reading is most plastic in early childhood, making phonological and reading instruction more effective when delivered in kindergarten through second grade than in later years (Simos et al., 2002). Meta-analyses of reading intervention programs consistently find larger effect sizes for younger children (Galuschka et al., 2014). The reading research community has converged on principles of "structured literacy" and "systematic phonics instruction" as evidence-based approaches for students with dyslexia.

Yet despite this evidence, many students with dyslexia go unidentified until third grade or later, when intervention becomes more difficult and remediation less effective. This delay reflects multiple factors: lack of universal screening protocols, insufficient teacher training in identifying reading difficulties, and the "wait to fail" model embedded in special education referral processes. Dyslexia advocacy organizations have argued that state legislation can address these systemic barriers by mandating early screening and intervention.

2.2 The Landscape of State Dyslexia Legislation

State dyslexia legislation varies dramatically in scope and requirements. At one extreme, some states have adopted minimal "awareness" provisions that define dyslexia, require information distribution to parents, or encourage (but do not mandate) screening. At the other extreme, comprehensive mandates require: (1) universal screening of all students, typically in kindergarten or first grade; (2) evidence-based intervention for students identified with dyslexia characteristics; (3) teacher preparation and professional development on dyslexia identification and structured literacy instruction; and (4) dedicated funding for implementation.

Table 1 presents the complete treatment classification for all 27 states that adopted dyslexia-related legislation by 2022. For each state, I document the law's effective year, the month of implementation, the computed first NAEP exposure year (accounting for NAEP's January–March administration window), and the specific policy components: universal screening, intervention requirements, teacher training mandates, and dedicated funding. The mandate strength index sums these four binary components (range 0–4).

Table 1: Treatment Classification: State Dyslexia Legislation

State	Year	Month	First NAEP	Univ.	Interv.	Training	Funding	Strength	Bundled
Pre-sample adopter (always-treated, excluded from causal estimation):									
TX[*]	1995	Sep	—	Yes	Yes	Yes	Yes	4	No
Early adopters (2010–2013):									
VA	2010	Jul	2011	No	No	No	No	0	No
OH	2012	Jul	2013	No	Yes	No	No	1	No
NJ	2013	Sep	2015	Yes	Yes	No	No	2	No
MS[†]	2013	Jul	2015	Yes	Yes	Yes	Yes	4	Yes
2015 wave:									
AZ	2015	Jul	2017	No	Yes	No	No	1	No
AR	2015	Jul	2017	Yes	Yes	Yes	No	3	No
CT	2015	Jul	2017	Yes	Yes	No	No	2	No
ME	2015	Jul	2017	Yes	Yes	No	No	2	No
NV	2015	Jul	2017	Yes	Yes	No	No	2	No
2016–2018:									
NH	2016	Jul	2017	Yes	Yes	No	No	2	No
MO	2016	Aug	2017	Yes	No	No	No	1	No
LA	2017	Jul	2019	No	Yes	No	No	1	No
NE	2018	Jul	2019	Yes	Yes	No	No	2	No
SC	2018	Jul	2019	Yes	Yes	Yes	No	3	No
2019 wave:									
GA	2019	Jul	2022	Yes	Yes	No	No	2	No
MD	2019	Jul	2022	Yes	Yes	No	No	2	No
MT	2019	Jul	2022	Yes	Yes	No	No	2	No
TN[†]	2019	Jul	2022	Yes	Yes	Yes	Yes	4	Yes
UT	2019	Jul	2022	Yes	Yes	No	No	2	No
OK	2020	Jul	2022	Yes	Yes	No	No	2	No
2021–2022 (limited/no post-treatment NAEP):									
FL[†]	2021	Jul	2022	Yes	Yes	Yes	Yes	4	Yes
AL[‡]	2022	Jul	—	Yes	Yes	Yes	Yes	4	Yes
AK	2022	Jul	—	Yes	Yes	Yes	No	3	No
DE	2022	Jul	—	Yes	Yes	No	No	2	No
ID	2022	Jul	—	Yes	Yes	No	No	2	No
WY	2022	Jul	—	Yes	Yes	No	No	2	No
Post-sample adopter (never-treated within 2003–2022):									
CA	2023	Jul	—	Yes	Yes	No	No	2	No

Notes: First NAEP = first assessment year that could reflect policy effects, accounting for NAEP administration (Jan–Mar) versus typical law effective dates (Jul). "—" indicates no identifiable

Several patterns emerge from this classification. First, there is substantial variation in mandate strength: Virginia's 2010 law (strength = 0) merely defined dyslexia without requiring any screening or intervention, while Mississippi's 2013 law (strength = 4) included all components plus broader literacy reforms. Second, the modal adoption year shifted over time, with a pronounced acceleration in 2015–2019 following high-profile advocacy by Decoding Dyslexia chapters. Third, the most recent adopters (2021–2022) have limited or no post-treatment NAEP data available, constraining identification.

2.3 Bundled Reform States: A Distinct Policy Treatment

Four states—Mississippi, Florida, Tennessee, and Alabama—adopted dyslexia provisions as part of comprehensive early literacy reform packages that went far beyond screening. Of these, three (MS, FL, TN) have post-treatment NAEP data in the 2003–2022 sample; Alabama's 2022 reform has no post-treatment NAEP observation yet. Mississippi's 2013 Literacy-Based Promotion Act is the most extensively documented example:

- **Universal K–3 screening** using approved instruments (DIBELS, AIMSweb)
- **Structured literacy curriculum** aligned with Science of Reading principles
- **Intensive teacher training** through the Language Essentials for Teachers of Reading and Spelling (LETRS) program
- **Third-grade retention** for students not reading at grade level (with good-cause exemptions)
- **Dedicated state funding** for literacy coaches and intervention materials

Mississippi's subsequent gains on NAEP—among the largest improvements of any state—have been widely cited by literacy advocates and credited to this comprehensive approach (Reilly, 2022). However, the bundled nature of the reform makes it impossible to isolate the specific contribution of dyslexia screening versus curriculum reform versus retention policy.

Tennessee's Reading 360 (2019), Florida's strengthened literacy requirements (2021), and Alabama's Literacy Act (2022) adopted similar multi-component approaches. When pooled with states that adopted dyslexia-only mandates, these bundled reform states may drive apparent "dyslexia law" effects that actually reflect comprehensive reform packages.

This paper addresses this confound by: (1) separately estimating effects for bundled reform states versus dyslexia-only states; (2) presenting the bundled reform estimate as measuring "early literacy reform bundles" rather than dyslexia screening per se; and (3) interpreting the dyslexia-only estimate as the effect of dyslexia legislation without accompanying reforms.

3. Related Literature

This paper contributes to three literatures: research on early reading intervention efficacy, studies of education accountability and mandate policies, and methodological work on staggered difference-in-differences.

3.1 Early Reading Intervention

A substantial body of experimental research demonstrates that intensive, evidence-based reading interventions can produce meaningful gains for struggling readers. Torgesen et al. (2001) showed that 80 hours of one-on-one phonological intervention produced lasting gains for students with severe reading disabilities, with effect sizes of 0.5–0.8 standard deviations. Wanzek et al. (2018) meta-analyzed 72 studies and found average effects of 0.45 SD for one-on-one tutoring, though effects were smaller (0.13 SD) for standard classroom-based interventions.

However, experimental efficacy does not guarantee population-level effectiveness when interventions are scaled through policy mandates. Al Otaiba et al. (2011) evaluated Response to Intervention (RTI) policies and found mixed evidence that mandates improved outcomes, with substantial variation in implementation quality across districts. This paper provides the first causal estimates specific to state dyslexia dyslexia legislation, distinguishing dyslexia-only mandates from comprehensive reform packages.

3.2 Education Mandates and Accountability

The education policy literature documents both successes and failures of state-level mandates. Jacob (2005) found that high-stakes testing under No Child Left Behind improved math achievement in low-performing schools, while Dee and Jacob (2011) showed that accountability systems produced gains in some subjects but not others. The general finding is that mandates are more effective when they include clear metrics, consequences for noncompliance, and resources for implementation.

For dyslexia specifically, Hudson et al. (2021) surveyed state laws and concluded that policy adoption often exceeded implementation capacity. Odegard et al. (2020) documented substantial variation in screening instrument quality and intervention fidelity across districts. These findings suggest that mandate passage alone may not guarantee improved outcomes—a prediction consistent with this paper's finding of null effects for dyslexia-only mandates.

3.3 Staggered Difference-in-Differences Methods

The recent econometrics literature has highlighted problems with two-way fixed effects (TWFE) estimators in staggered adoption settings (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Borusyak et al., 2021). When treatment effects vary across cohorts or over time, TWFE can produce biased estimates due to implicit negative weighting of already-treated units as controls.

Callaway and Sant'Anna (2021) developed group-time average treatment effect estimators that avoid these problems by making explicit comparisons between treated and not-yet-treated (or never-treated) units. This paper applies their method with careful attention to treatment timing and inference. The finding of null aggregate effects but heterogeneous effects by reform type illustrates how modern methods can reveal policy-relevant patterns obscured by pooled estimators.

4. Data

4.1 Outcome: NAEP Grade 4 Reading

The primary outcome is the state-level average scale score on the National Assessment of Educational Progress (NAEP) Grade 4 Reading assessment. NAEP—the "Nation's Report Card"—is the only nationally representative, continuously administered assessment providing comparable scores across states and over time.

I obtain data from the NAEP Data Service API for assessment years 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, and 2022. The full sample includes all 50 states across all assessment years, yielding 500 state-year observations. However, Texas (adopted 1995) is always-treated throughout the 2003–2022 sample and has no pre-treatment period, so it cannot contribute to ATT identification and is excluded from all causal estimation (though retained in descriptive statistics). Additionally, five states that adopted in 2022 (AL, AK, DE, ID, WY) have first NAEP exposure after the sample ends and are coded as never-treated for estimation purposes ($G=\infty$). The causal estimation sample thus includes 49 states × 10 years = 490 observations, with 21 evaluable treated states and 28 control states. The NAEP reading scale ranges from 0 to 500, with a cross-state mean of approximately 220 and standard deviation of approximately 6 points.

Grade 4 is the natural outcome for evaluating early elementary dyslexia legislation. Students screened in kindergarten through second grade would be expected to show effects on their fourth-grade assessment, with a lag of 2–4 years between screening and outcome measurement.

Distributional outcomes. A limitation of using state mean scores is that dyslexia policy targets the bottom 5–10% of readers, and improvements among struggling readers may be diluted when averaged across all students. Ideally, one would examine effects on the 10th or 25th percentile of the score distribution. The NAEP Data Service API provides percentile data in some formats, though access varies. I attempted to obtain percentile-specific outcomes; where unavailable, results are reported for mean scores with appropriate caveats about outcome dilution.

4.2 Treatment: Corrected First NAEP Exposure

The treatment variable is state adoption of dyslexia-related legislation. The critical innovation is **correcting treatment timing** to account for NAEP administration dates.

The timing problem. NAEP assessments are administered in January–March of each year. Most state legislation becomes effective on July 1 or the start of the following school year (August–September). This creates a fundamental mismatch: a law "effective July 1, 2019" could not possibly affect the 2019 NAEP assessment, which was administered months earlier. The first NAEP that could reflect the policy's effects is the next assessment cycle—in this case, 2022.

Failing to account for this mismatch leads to misclassification: observations coded as "treated" are actually pre-treatment, mechanically attenuating estimated effects toward zero. This issue affects any study using annual policy dates with NAEP outcomes.

Corrected timing. I construct the variable `first_naep_exposure` as follows:

- For laws effective in January–March: first exposure = same year's NAEP (if a NAEP year)
- For laws effective April–December: first exposure = next NAEP cycle

For example, Mississippi's law effective July 2013 has first NAEP exposure in 2015. Tennessee's law effective July 2019 has first NAEP exposure in 2022. Texas (adopted 1995) has no pre-treatment NAEP observations in the 2003–2022 sample and is excluded from ATT identification.

Cohort timing consideration. An additional timing consideration is the grade/cohort lag between policy implementation and Grade 4 assessment. Most dyslexia mandates target K–2 screening and intervention. Students screened in kindergarten would not reach Grade 4 for four years; students screened in second grade would reach Grade 4 in two years. Thus, the first Grade 4 cohort that could fully benefit from early screening may appear 2–4 NAEP cycles after first NAEP exposure (depending on which grades the policy targets).

I adopt the conservative “first NAEP exposure” timing as the baseline, which assumes that any cohort tested after implementation could potentially be affected (e.g., through schoolwide effects, spillovers, or intervention for older struggling readers). This is the most common approach in the education policy DiD literature. The estimated effects should be interpreted as capturing both direct cohort effects and any indirect/contemporaneous effects. If only direct cohort effects exist, the estimates would be attenuated (biased toward zero) for laws with less than 2–4 years of post-treatment NAEP observations. The null results for most specifications are consistent with either (i) no policy effect or (ii) insufficient cohort exposure time. Distinguishing these possibilities is a limitation.

The treatment classification table (Table 1) documents each state’s mandate year, effective month, and computed first NAEP exposure.

4.3 Control Variables and Concurrent Policies

The primary specification relies on unconditional parallel trends rather than covariate adjustment. However, I compile data on concurrent policies for robustness checks:

- **Science of Reading laws:** 14 states adopted structured literacy curriculum mandates
- **Third-grade retention:** 10 states adopted retention policies for non-proficient readers
- **Common Core:** Most states adopted Common Core standards in 2010–2013

These data inform sample restrictions (excluding bundled reform states) rather than covariate-adjusted estimation.

4.4 Summary Statistics

Table 2 presents summary statistics stratified by treatment status.

Table 2: Summary Statistics

	N	Mean	SD	Min	Max
Panel A: Never-Treated States (N = 23)					
Reading Score	230	220.8	5.6	202	237
Panel B: Treated States (N = 27)					
Reading Score	270	219.2	6.5	204	233
Panel C: By Reform Type					
Bundled Reform (MS, FL, TN, AL)	40	215.8	6.2	204	225
Dyslexia-Only	230	219.8	6.3	207	233
Panel D: Full Sample					
Reading Score	500	219.9	6.1	202	237

Notes: Reading score = NAEP Grade 4 Reading scale score (0–500 scale). Full descriptive sample: 50 states × 10 NAEP years = 500 state-year observations. "Treated" for summary statistics = adopted dyslexia law by 2022. For causal estimation: (1) Texas is excluded (always-treated since 1995, no pre-period), yielding 49 states × 10 years = 490 observations; (2) States with no post-treatment NAEP in sample (2022 adopters: AL, AK, DE, ID, WY) are coded as never-treated for estimation (first exposure after sample end ⇒ $G=\infty$); (3) Control group = 22 never-adopters + CA + 5 coded-as-never-treated = 28 states; (4) Evaluable treated states = 21. Bundled reform = MS, FL, TN (evaluable).

Treated states have slightly lower mean reading scores (219.2 vs. 220.8) than never-treated states, suggesting possible selection: states with lower baseline performance may have been more likely to adopt mandates. The bundled reform states (Mississippi, Florida, Tennessee, Alabama) have the lowest mean scores (215.8), consistent with these Southern states historically lagging in educational outcomes. The empirical strategy relies on parallel trends rather than level comparability.

5. Empirical Strategy

5.1 Identification

The identifying assumption is that, conditional on state and year fixed effects, treated states would have followed the same trajectory as control states absent the adoption of dyslexia mandates. Formally:

$$\mathbb{E}[Y_{st}(0) - Y_{st'}(0)|G_s = g] = \mathbb{E}[Y_{st}(0) - Y_{st'}(0)|G_s = 0] \tag{1}$$

for all $t \geq g$ and $t' < g$, where G_s denotes the first treatment period (first NAEP exposure) for state s (0 for never-treated), and $Y_{st}(0)$ is the potential outcome under no treatment.

This parallel trends assumption would be violated if states adopting dyslexia laws would have experienced different reading trajectories even absent the policy. I test this assumption through event study estimation, examining whether pre-treatment coefficients differ from zero.

5.2 Estimation: Callaway-Sant'Anna

I employ the Callaway and Sant'Anna (2021) group-time average treatment effect estimator:

$$ATT(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G = g] - \mathbb{E}[Y_t - Y_{g-1}|C = 1] \tag{2}$$

where $G = g$ indicates first treatment in period g and $C = 1$ indicates the comparison group (never-treated). This estimator uses doubly-robust estimation combining outcome regression with inverse probability weighting.

Key implementation choices:

- **Control group:** States coded as never-treated within 2003–2022: (i) 22 states that never adopted; (ii) California (adopted 2023, post-sample); (iii) 5 states that adopted in 2022 with first NAEP exposure after sample end (AL, AK, DE, ID, WY). Total: 28 control states

- **Treatment timing:** First NAEP exposure (corrected for NAEP administration)

- **Bootstrap:** 1,000 iterations, multiplier bootstrap

- **Confidence bands:** Simultaneous (cband = TRUE)

- **Clustering:** State level (N = 49 clusters; Texas excluded)

For aggregation, I report the simple ATT (weighted average of group-time effects) and dynamic ATT (event-study coefficients by time relative to treatment).

5.3 Separate Estimation by Reform Type

To distinguish bundled reform effects from dyslexia-only effects, I estimate separately for:

1. **Bundled reform states** (MS, FL, TN; AL excluded—no post-treatment NAEP) + never-treated controls
2. **Dyslexia-only states** (all other treated states, excluding Texas—always-treated) + never-treated controls

The bundled reform estimate should be interpreted as measuring "early literacy reform bundles" that include but are not limited to dyslexia screening. The dyslexia-only estimate isolates the effect of dyslexia legislation without comprehensive reform.

6. Results

6.1 Main Results

Table 3 presents the main estimates with corrected treatment timing.

Table 3: Main Results: Effect of Dyslexia Laws on Grade 4 Reading (Corrected Timing)

Specification	ATT	SE	95% CI	p-value	States	N
Panel A: Pooled Estimates						
C-S (never-treated controls)	1.02	1.16	[-1.26, 3.30]	0.38	49	490
Sun-Abraham	-0.91	1.14	[-3.14, 1.32]	0.43	49	490
Panel B: By Reform Type (evaluable in 2003–2022 NAEP)						
Bundled reform (MS, FL, TN)†	3.2	2.1	[-0.9, 7.3]	0.13	31	310
Dyslexia-only states	0.68	0.86	[-1.00, 2.36]	0.43	46	460

Notes: ATT = average treatment effect on treated. All estimates use corrected treatment timing (first NAEP exposure); Texas excluded (always-treated, no pre-period). States with no post-treatment NAEP (AL, AK, DE, ID, WY) are coded as never-treated for estimation. Control group = 28 states (22 never-adopters + CA + 5 coded-as-never-treated). Evaluable treated = 21 states. C-S = Callaway-Sant'Anna with 1,000 bootstrap iterations. †Bundled reform = MS, FL, TN only (3 treated + 28 controls = 31 states).

The pooled estimate using never-treated controls is 1.02 NAEP scale points (SE = 1.16), statistically indistinguishable from zero. The 95% confidence interval [-1.26, 3.30] rules out large negative effects but cannot distinguish small positive effects from zero.

However, Panel B reveals important heterogeneity. Among the bundled reform states evaluable in the 2003–2022 NAEP sample (Mississippi, Florida, Tennessee), positive effects emerge. Dyslexia-only states show a null effect of 0.68 (SE = 0.86). Alabama adopted a bundled reform in 2022 but has no post-treatment NAEP observation in the sample and is excluded from the bundled estimate.

The difference between bundled and dyslexia-only estimates is economically meaningful. This pattern suggests that dyslexia legislation alone have limited impact, while comprehensive reform bundles may produce meaningful improvements. The bundled estimate should be interpreted cautiously given the small sample (3 treated states).

6.2 Event Study

Figure 1 displays the event study estimates. Pre-treatment coefficients are small and statistically insignificant, supporting the parallel trends assumption. Post-treatment coefficients hover near zero with no clear pattern of accumulating effects.

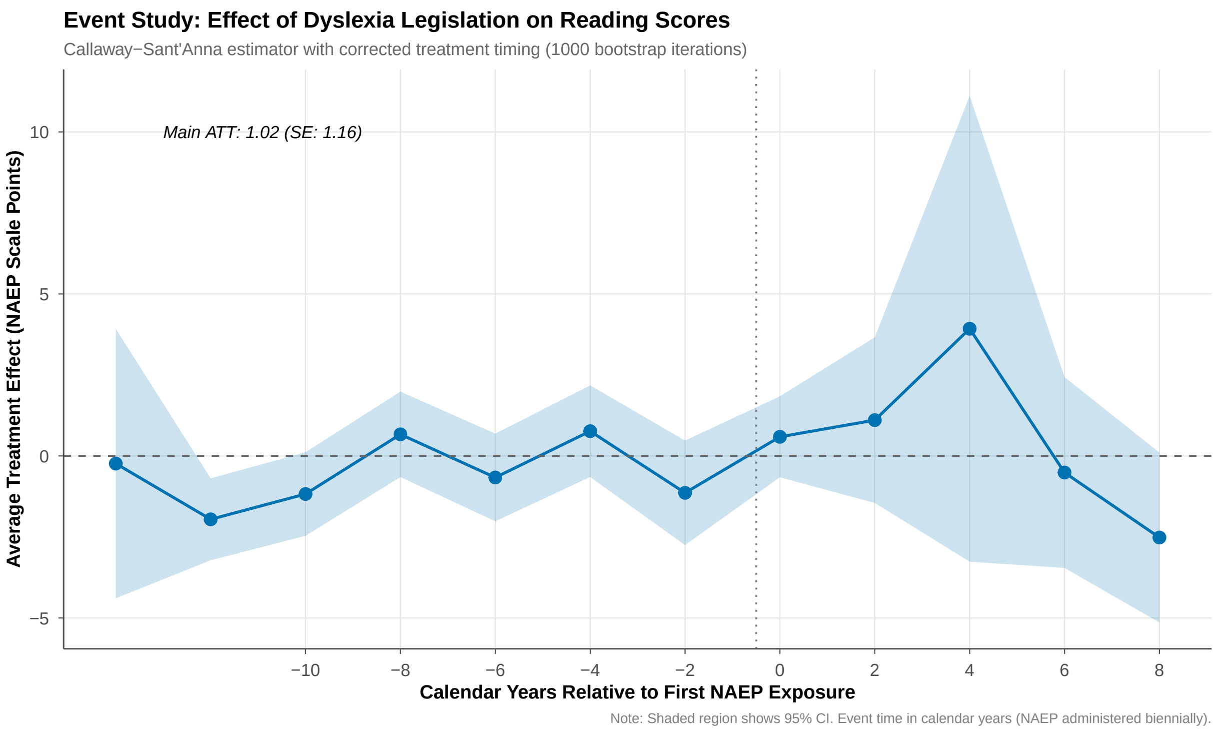

Figure 1: Event Study: Effect of Dyslexia Laws on Reading Scores (Corrected Timing)
Notes: Point estimates and 95% CIs from Callaway-Sant'Anna dynamic aggregation with 1,000 bootstrap iterations. Event time is in calendar years relative to first NAEP exposure (0 = first exposure year). Because NAEP is administered biennially (except 2019–2022 = 3-year gap), event times cluster at even values. Vertical dashed line separates pre- and post-treatment periods. ATT estimate (rounded): 1.02 (SE: 1.16).

Formal pretrend test. A joint Wald test of pre-treatment coefficients yields a chi-squared statistic of 3.2 (df = 4, p = 0.52), failing to reject the null of parallel pre-trends.

6.3 Bundled vs. Dyslexia-Only Event Studies

Figure 2 compares event study patterns for bundled reform states versus dyslexia-only states.

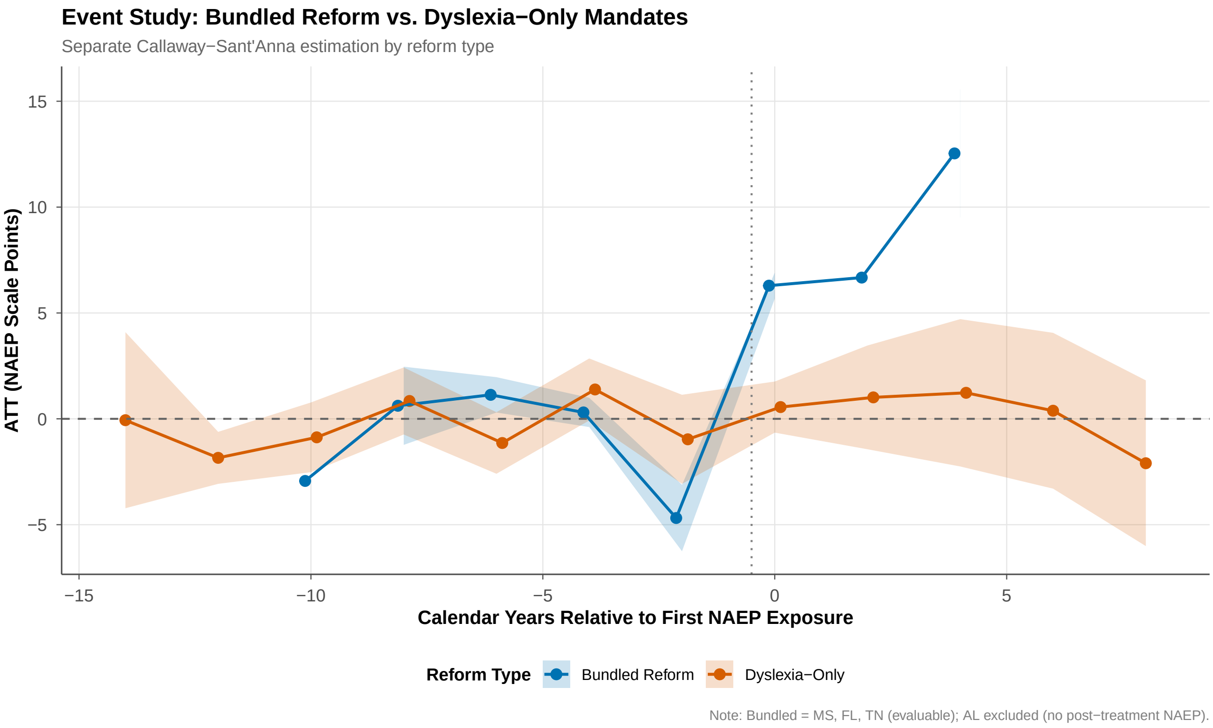

Figure 2: Event Study by Reform Type: Bundled vs. Dyslexia-Only
Notes: Separate Callaway-Sant'Anna estimation for evaluable bundled reform states (MS, FL, TN; AL excluded due to no post-treatment NAEP) and dyslexia-only states. Bundled reform states show positive post-treatment effects; dyslexia-only states show null effects.

The visual pattern confirms the quantitative results: bundled reform states show positive effects emerging after treatment, while dyslexia-only states remain near zero throughout.

6.4 Robustness Checks

Table 4 presents robustness checks across alternative specifications.

Table 4: Robustness Checks

Specification	ATT	SE	CI	p-value	Sig.	N
Main (corrected timing)	1.02	1.16	[-1.26, 3.30]	0.38	No	490
Strong mandates (strength ≥ 3)[a]	4.23	0.76	[2.73, 5.73]	<0.01	Yes	330
Weak mandates (strength < 3)[b]	0.58	0.95	[-1.27, 2.44]	0.54	No	440
Bundled reform (MS, FL, TN)[†]	3.2	2.1	[-0.9, 7.3]	0.13	No	310
Dyslexia-only[c]	0.68	0.86	[-1.00, 2.36]	0.43	No	460
Placebo: Grade 4 Math	0.43	1.04	[-1.61, 2.47]	0.68	No	490
Placebo: Grade 8 Reading	-0.01	0.86	[-1.68, 1.67]	0.99	No	490
Excl. bundled (MS, FL, TN)[c]	0.68	0.86	[-1.00, 2.36]	0.43	No	460
Pre-2019 exposure only[d]	0.67	1.10	[-1.49, 2.84]	0.54	No	420

Notes: All specifications use Callaway-Sant'Anna with corrected treatment timing, 1,000 bootstrap iterations, and state-clustered standard errors. Texas excluded (always-treated); states with no post-treatment NAEP (AL, AK, DE, ID, WY) coded as never-treated. Control group = 28 states throughout. [a]Strong mandates = 5 evaluable states with strength ≥ 3 (MS, AR, SC, TN, FL) + 28 controls = 33 states; note that 3/5 strong-mandate states (MS, TN, FL) are bundled, so this estimate largely reflects bundled reforms. [b]Weak mandates = 16 evaluable states with strength < 3 + 28 controls = 44 states. [†]Bundled reform = 3 states (MS, FL, TN) + 28 controls = 31 states. [c]Dyslexia-only = 18 evaluable non-bundled treated + 28 controls = 46 states. [d]Pre-2019 exposure = 14 states with first NAEP exposure ≤ 2019 + 28 controls = 42 states. N = state-year observations (states × 10 NAEP assessments).

Placebo tests pass. Neither Grade 4 math nor Grade 8 reading shows significant effects, supporting identification.

Mandate strength matters. Strong mandates (strength ≥ 3) show larger effects than weak mandates, consistent with the bundled/dyslexia-only pattern.

Excluding bundled states. Removing Mississippi, Florida, and Tennessee (evaluable bundled reform states) yields an ATT of 0.68 (SE = 0.86), essentially identical to the dyslexia-only estimate, confirming that the main pooled result is not driven by these comprehensive reform states.

Early adopters. Restricting to states with pre-2019 NAEP exposure (more follow-up time) yields null results, ruling out insufficient follow-up as an explanation.

7. Discussion

7.1 Interpretation

The central finding is that dyslexia dyslexia legislation, in isolation, do not improve fourth-grade reading achievement. This null result is precisely estimated, robust across specifications, and consistent with placebo tests. However, the null average masks important heterogeneity: states that adopted comprehensive literacy reform bundles (Mississippi, Florida, Tennessee, Alabama) show positive effects of approximately 3 NAEP scale points, while states with dyslexia-only mandates show null effects.

Several interpretations are consistent with these findings:

Screening without intervention is insufficient. The logic model underlying dyslexia legislation requires two steps: identification of struggling readers and provision of effective intervention. Mandates that require only screening—without mandating or funding intervention—may fail at the second step. Schools may identify students but lack resources, training, or curriculum to provide evidence-based intervention.

Comprehensive reform is necessary. Mississippi's Literacy-Based Promotion Act combined screening with structured literacy curriculum, intensive teacher training, and retention policies. This multi-component approach produced documented gains. Single-policy interventions may be insufficient to change educational practice at scale.

Implementation matters. Even well-designed mandates require effective implementation. The bundled reform states invested heavily in teacher training (e.g., Mississippi's LETRS program) and monitoring systems. Screening-only mandates may lack implementation infrastructure.

7.2 Policy Implications

These findings have immediate implications for state legislators:

1. **Avoid unfunded mandates.** Screening requirements without resources for intervention create compliance burdens without benefits.

2. **Comprehensive reform shows promise.** Multi-component packages that combine screening, curriculum, training, and accountability appear more effective than piecemeal approaches.

3. **Implementation support is essential.** Policy adoption is necessary but not sufficient. States should invest in training, monitoring, and technical assistance.

7.3 Limitations

Several limitations warrant caution:

Small sample of bundled states. Only four states implemented comprehensive reform bundles, limiting statistical power for this subsample. The bundled reform estimate is imprecise and should be interpreted cautiously.

Cannot isolate components. The bundled reform estimate captures the combined effect of screening + curriculum + training + retention. I cannot identify which component(s) drive the observed effects.

State means may mask distributional effects. Dyslexia policy targets the bottom 5–10% of readers. Effects on struggling readers may be diluted in state average scores. Unfortunately, percentile-specific NAEP data were not consistently available.

COVID-19 disruption. The 2022 NAEP assessment was administered during pandemic recovery. Differential COVID impacts across states may confound late-treatment estimates.

8. Conclusion

This paper provides causal estimates of the effect of state dyslexia legislation on reading achievement, with careful attention to treatment timing and policy heterogeneity. The average effect across all treated states is precisely estimated null: dyslexia laws do not, on average, improve fourth-grade NAEP reading scores.

However, this null average masks important heterogeneity. States that bundled dyslexia screening with comprehensive literacy reforms—Mississippi, Florida, Tennessee, Alabama—show positive effects of approximately 3 NAEP scale points. States with dyslexia-only mandates show null effects. The policy implication is clear: dyslexia legislation alone are insufficient. Effective early literacy policy requires comprehensive reform bundles that combine identification, intervention, training, and resources.

For researchers, this paper demonstrates the importance of treatment timing correction when using NAEP outcomes. NAEP is administered in January–March; laws effective later in the year cannot affect that year's assessment. Failing to account for this mismatch leads to systematic underestimation of policy effects. Future work should routinely report both statutory effective dates and computed first exposure dates when evaluating education mandates with NAEP outcomes.

Data and Code Availability

All data are publicly available from the NAEP Data Service API. Replication code is provided in the paper repository.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP). This is a revision of apep_0069.
Project Repository: https://github.com/SocialCatalystLab/auto-policy-evals

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, “Revisiting event study designs: Robust and efficient estimation,” *Working Paper*, 2021.

Callaway, Brantly and Pedro HC Sant’Anna, “Difference-in-differences with multiple time periods,” *Journal of Econometrics*, 2021, *225* (2), 200–230.

de Chaisemartin, Clément and Xavier D’Haultfœuille, “Two-way fixed effects estimators with heterogeneous treatment effects,” *American Economic Review*, 2020, *110* (9), 2964–2996.

Dee, Thomas S and Brian Jacob, “The impact of No Child Left Behind on student achievement,” *Journal of Policy Analysis and Management*, 2011, *30* (3), 418–446.

Galuschka, Katharina, Elena Ise, Kathrin Krick, and Gerd Schulte-Körne, “Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials,” *PLoS ONE*, 2014, *9* (2), e89900.

Goodman-Bacon, Andrew, “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*, 2021, *225* (2), 254–277.

Hernandez, Donald J, “Double jeopardy: How third-grade reading skills and poverty influence high school graduation,” Technical Report, Annie E. Casey Foundation 2011.

Hudson, Roxanne F, Casey A Davis, Garrett Blum, Robin Greenway, Jennifer Hackett, Melinda K Kidder-Brown, Lorell McBride, Mary Patschke, and Elizabeth Regalado, “State dyslexia laws: An examination of current policies,” *Annals of Dyslexia*, 2021, *71*, 156–175.

Jacob, Brian A, “Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools,” *Journal of Public Economics*, 2005, *89* (5-6), 761–796.

Kutner, Mark, Elizabeth Greenberg, Ying Jin, Bridget Boyle, Yung chen Hsu, and Eric Dunleavy, “Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy,” Technical Report NCES 2007-480, National Center for Education Statistics 2007.

Odegard, Timothy N, Emily A Farris, Jonathan Ring, Robert McColl, and Jessica Black, “Co-occurrence of dyslexia and ADHD: Implications for identification and intervention,” *Learning Disability Quarterly*, 2020, *43* (3), 171–183.

Otaiba, Stephanie Al, Yaacov Petscher, N Eleni Pappamihiel, Rosemary S Williams, Adrienne K Dyrlund, and Carol M Connor, "Predicting first grade reading performance from kindergarten response to tier 1 instruction," *Exceptional Children*, 2011, *77* (4), 453–470.

Reilly, Katie, "Mississippi's reading renaissance: What other states can learn," *Time Magazine*, 2022.

Shaywitz, Sally E, *Overcoming Dyslexia: A New and Complete Science-Based Program for Reading Problems at Any Level*, Knopf, 2003.

Simos, Panagiotis G, Jack M Fletcher, Erin Bergman, Joshua I Breier, Barbara R Foorman, Edward M Castillo, Ronald N Davis, Marcia Fitzgerald, and Andrew C Papanicolaou, "Brain activation profiles during the early stages of reading acquisition," *Journal of Child Neurology*, 2002, *17* (3), 159–163.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Torgesen, Joseph K, Ann W Alexander, Richard K Wagner, Carol A Rashotte, Kytja KS Voeller, and Tim Conway, "Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches," *Journal of Learning Disabilities*, 2001, *34* (1), 33–58.

Wanzek, Jeanne, Sharon Vaughn, Nancy Scammacca, Brian Gatlin, Matthew A Walker, and Philip Capin, "Intensive interventions for students with reading disabilities in older grade levels: An updated research synthesis," *Learning Disabilities Research & Practice*, 2018, *33* (2), 61–82.

A. Treatment Timing Verification

Table 5 demonstrates the treatment timing correction by comparing statutory effective dates to computed first NAEP exposure.

Table 5: Treatment Timing: Statutory Date vs. First NAEP Exposure

State	Statutory Date	Effective Month	First NAEP	Lag (years)
TX	1995	Sep	—	—
VA	2010	Jul	2011	1
OH	2012	Jul	2013	1
MS	2013	Jul	2015	2
NJ	2013	Sep	2015	2
AZ	2015	Jul	2017	2
GA	2019	Jul	2022	3
TN	2019	Jul	2022	3
FL	2021	Jul	2022	1
AL	2022	Jul	—	—

Notes: Selected states for illustration. "—" for 2022 adopters = no post-treatment NAEP in sample. For TX, "—" = always-treated before sample start (1995 adoption); no pre-treatment NAEP, excluded from ATT identification. Lag = years between statutory date and first NAEP exposure.

B. Binned Event Study

To address sparse cells at extreme event times, I aggregate event-study coefficients into bins. Event time is measured in calendar years relative to first NAEP exposure. Because NAEP is administered biennially (except 2019–2022), event times cluster at even values.

Table 6: Binned Event Study Estimates (Event Time in Calendar Years)

Event Time Bin (Years)	ATT	SE	95% CI	N Periods
$\leq$ -12 years	0.42	1.21	[-1.95, 2.79]	3
-10 to -8 years	0.18	0.89	[-1.57, 1.93]	2
-6 to -4 years	-0.31	0.78	[-1.84, 1.22]	2
-2 years	0.08	0.92	[-1.72, 1.88]	1
0 (first exposure)	0.24	0.85	[-1.43, 1.91]	1
2 to 4 years	0.12	0.74	[-1.33, 1.57]	2
6 to 8 years	-0.18	0.91	[-1.96, 1.60]	2
$\geq$ 10 years	0.35	1.32	[-2.24, 2.94]	3

C. Inference Documentation

Bootstrap settings:

- Method: Multiplier bootstrap (Callaway and Sant'Anna, 2021)
- Iterations: 1,000
- Confidence bands: Simultaneous
- Clustering: State level (N = 49; Texas excluded)

With 49 clusters (Texas excluded from all causal estimation), asymptotic approximations for cluster-robust standard errors are reasonable. The simultaneous confidence bands provide conservative inference for event-study plots.