Legislating the Schoolyard Online: Do Anti-Cyberbullying Laws Reduce Youth Suicide Risk?

APEP Autonomous Research* @ai1scl

February 10, 2026

Abstract

Between 2006 and 2015, 48 U.S. states adopted laws requiring schools to address cyberbullying—yet youth suicide rates continued rising throughout this period. I exploit the staggered timing of state anti-cyberbullying legislation to estimate its causal effect on adolescent mental health using the Youth Risk Behavior Surveillance System (1991–2017). Employing both Sun and Abraham (2021) heterogeneity-robust difference-in-differences and standard two-way fixed effects, I find that anti-cyberbullying laws had no statistically significant effect on suicide ideation, suicide attempts, or depressive symptoms among high school students. The TWFE estimate for suicide ideation is 0.111 percentage points (SE = 0.457, $p = 0.81$); for depression, -0.202 (SE = 0.423, $p = 0.63$). The one borderline-significant Sun-Abraham estimate—for suicide attempts (1.170 pp, $p = 0.047$)—is in the *wrong* direction (an increase) and vanishes under randomization inference ($p = 0.26$), consistent with a false positive. The null extends to states with criminal penalties for cyberbullying and persists across sex, suggesting that legislative mandates—whether through school policy requirements or criminal sanctions—are insufficient to meaningfully reduce the mental health burden of online harassment among adolescents. These findings inform the current debate over social media regulation by demonstrating that first-generation anti-cyberbullying statutes did not deliver measurable mental health benefits, underscoring the need for more targeted interventions.

JEL Codes: I18, I28, K42, J13

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch

Keywords: cyberbullying, youth mental health, suicide, social media regulation, difference-in-differences, staggered adoption

1. Introduction

In the spring of 2017, a high school student in Florida took her own life after months of relentless online harassment. Her case joined a growing roster of adolescent suicides linked to cyberbullying that had already prompted legislative action in nearly every state. Yet even as state legislatures passed anti-cyberbullying laws, youth suicide rates rose steadily—climbing 56% among 10–24 year-olds between 2007 and 2017 (Curtin, 2020). This paper asks whether these laws made any difference at all.

The question sits at the intersection of two urgent policy debates. First, mounting evidence links social media use to adolescent mental health deterioration (Twenge et al., 2018; Haidt, 2023), prompting calls for platform regulation, age verification mandates, and bans on social media for minors. Second, policymakers and parents desperately seek evidence on what actually works—a question that the existing literature, dominated by cross-sectional correlations between cyberbullying victimization and mental health outcomes, cannot credibly answer (John et al., 2018). Before investing in the next generation of social media regulations, we should understand whether the first generation accomplished its goals.

I exploit the staggered adoption of anti-cyberbullying laws across 48 U.S. states between 2006 and 2015 to estimate their causal effect on youth mental health. These laws—which variously mandate school anti-cyberbullying policies, require reporting mechanisms, and impose criminal penalties for electronic harassment—provide a natural experiment. States adopted them at different times for idiosyncratic reasons (high-profile bullying cases, legislative session timing, advocacy group pressure), creating the variation needed for credible identification.

My primary data source is the Youth Risk Behavior Surveillance System (YRBS), a biennial CDC survey of high school students that has tracked suicide ideation, suicide attempts, and depressive symptoms at the state level since the early 1990s. Crucially, the YRBS also measures electronic bullying victimization from 2011 onward, allowing me to test the first-stage hypothesis that these laws actually reduced cyberbullying. I combine 14 biennial waves (1991–2017) covering approximately 40 states per wave with a hand-coded treatment matrix documenting each state's anti-cyberbullying law adoption year and provisions.

I employ two complementary estimation strategies. My primary approach uses the Sun and Abraham (2021) heterogeneity-robust estimator, which is designed for staggered adoption settings where treatment effects may vary across cohorts and time—exactly the concern raised by Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœuille (2020) about standard two-way fixed effects (TWFE). I supplement this with conventional TWFE as a benchmark and attempt Callaway and Sant'Anna (2021) estimation on a restricted sample. All specifications include state and year fixed effects with standard errors clustered at the

state level.

The results tell a clear story: anti-cyberbullying laws failed to move the needle on youth suicide ideation, suicide attempts, depression, or—on the first-stage—electronic bullying victimization itself. The TWFE point estimate for suicide ideation is 0.111 percentage points (SE = 0.457, $p = 0.81$), with a 95% confidence interval spanning roughly ±1 percentage point around zero (against a baseline rate of 17.5%). The Sun-Abraham estimate is similarly null at 0.792 (SE = 1.437, $p = 0.58$). The one borderline result—a Sun-Abraham estimate of 1.170 for suicide attempts ($p = 0.047$)—points in the wrong direction—it implies that the laws *increased* suicide attempts, though this is almost certainly a statistical artifact—and is not robust to randomization inference ($p = 0.26$). The null persists across every specification I examine: Sun-Abraham heterogeneity-robust estimation, Bacon decomposition diagnostics, randomization inference, alternative treatment timing windows, and disaggregation by sex and law type (criminal penalties vs. school-policy-only mandates).

Three dimensions of heterogeneity are particularly informative. First, the null holds separately for females—who experience cyberbullying at substantially higher rates and suffer more severe mental health consequences—with point estimates for suicide ideation (0.736, $p = 0.19$) and depression (−0.506, $p = 0.37$) both insignificant, suggesting the laws failed even among the population most likely to benefit. Second, states that enacted criminal penalties for cyberbullying show no stronger effects than states that merely mandated school policies (e.g., suicide ideation: criminal 0.430 vs. school −0.027, both insignificant), casting doubt on the deterrence channel. Third, the event study reveals no dynamic pattern: neither gradual improvement as laws take hold nor sudden improvement upon adoption.

This paper contributes to three literatures. First, it advances the literature on cyberbullying and youth mental health by providing the first estimates using heterogeneity-robust staggered DiD methods and actual mortality-linked outcome data spanning 26 years. The closest prior work is Nikolaou (2017), who uses YRBS data with cyberbullying laws as instruments in a bivariate probit framework and finds that cyberbullying increases suicidal behaviors. My approach differs in using modern DiD methods designed for staggered adoption, examining a much longer time series, and focusing on the direct policy question of whether laws—not cyberbullying itself—affect outcomes.

Second, I contribute to the growing literature evaluating technology-related regulations, including studies of data privacy laws (Miller and Tucker, 2011; Goldfarb and Tucker, 2011) and content moderation policies. My null finding adds to a pattern in this literature: legislative responses to technology-mediated harms often fail to produce measurable improvements, perhaps because the technology evolves faster than the regulatory apparatus.

Third, the paper speaks to the broader policy evaluation literature on "symbolic legislation"—

laws passed in response to moral panics that may serve expressive rather than instrumental functions (Ben-Shahar and Schneider, 2021). Anti-cyberbullying laws were adopted rapidly in the wake of high-profile tragedies, often with minimal attention to enforcement mechanisms or evidence-based design. The null finding is consistent with the hypothesis that these laws were primarily expressive acts.

2. Institutional Background and Policy Setting

2.1 The Rise of Cyberbullying as a Policy Concern

The emergence of cyberbullying as a distinct social phenomenon tracked the rapid adoption of social media platforms among American teenagers in the mid-2000s. MySpace launched in 2003, Facebook opened to high school students in 2005, and by 2007 approximately 55% of online teens used social networking sites (Lenhart and Madden, 2007). With these platforms came a new form of peer aggression: electronic harassment that could follow victims home from school, reach vast audiences instantly, and persist indefinitely online.

Several high-profile cases galvanized public attention and legislative action. The 2006 suicide of Megan Meier, a 13-year-old Missouri girl cyberbullied through a fake MySpace profile created by an adult neighbor, sparked national outrage and directly motivated legislative proposals in multiple states. In 2010, the suicide of Tyler Clementi, a Rutgers University student whose roommate secretly streamed his intimate encounter online, further intensified legislative pressure. These cases—and dozens of less publicized ones—created political conditions in which passing anti-cyberbullying legislation became nearly obligatory for state legislators.

2.2 The Legislative Landscape

Between 2006 and 2015, 48 states adopted laws that explicitly address cyberbullying or electronic harassment within their anti-bullying statutes. The legislative timeline reflects several distinct phases. In the *pioneer phase* (2006–2007), Idaho and South Carolina became the first movers in 2006, amending existing bullying statutes to include "electronic" or "technological" harassment. Six additional states followed in 2007 (Iowa, Minnesota, Missouri, Ohio, Oregon, Washington), often in direct response to the national attention generated by the Megan Meier case and advocacy by groups like the Cyberbullying Research Center. Missouri's law, enacted in Meier's home state, was explicitly named in reference to her case and became a template for subsequent legislation.

The *rapid diffusion phase* (2008–2012) saw the bulk of state action. Six states adopted

in 2008, five in 2009, four in 2010, eight in 2011, and ten in 2012—the peak adoption year. This wave was catalyzed in part by federal attention: in 2010, the U.S. Department of Education held the first Federal Bullying Prevention Summit, and the Obama administration launched StopBullying.gov, a federal clearinghouse that explicitly urged states to adopt anti-cyberbullying provisions. While no federal anti-cyberbullying statute was enacted, the federal push created strong political incentives for state legislators to act. By this period, passing an anti-cyberbullying bill had become a low-cost, high-visibility signal of responsiveness to a salient public concern.

In the *late adopter phase* (2013–2015), the remaining states closed the gap: Indiana and Utah in 2013, with Montana completing the wave in 2015. Only Alaska and Wisconsin never specifically included cyberbullying provisions in their anti-bullying laws during the study period, though both have general anti-bullying statutes that could in principle be applied to electronic harassment. The near-universality of adoption—48 of 50 states within a decade—is itself notable and characteristic of "policy diffusion" dynamics in which interstate competition and advocacy pressure produce rapid convergence.

The laws vary considerably in their provisions, but generally fall into two categories that represent fundamentally different theories of behavior change. *School policy mandates* (37 states) require school districts to adopt anti-bullying policies that specifically address electronic harassment. These mandates typically include several interlocking components. First, *policy adoption requirements*: districts must draft and publicize written anti-bullying policies that explicitly enumerate cyberbullying and electronic harassment as prohibited conduct, often specifying platforms (text messaging, social media, email) by name. Second, *reporting mechanisms*: states require schools to establish formal procedures through which students, parents, and staff can report cyberbullying incidents, frequently mandating anonymous reporting options and designating specific school personnel (typically a counselor, assistant principal, or "bullying prevention coordinator") as the responsible official. Third, *investigation procedures*: upon receiving a report, schools must conduct a documented investigation within a specified timeframe—commonly 5 to 10 school days—including interviews with the alleged victim, perpetrator, and witnesses, and must notify parents of all parties involved. Fourth, *disciplinary actions*: substantiated cyberbullying must trigger graduated consequences ranging from counseling and behavioral contracts through suspension to expulsion in severe cases, with many states requiring that disciplinary responses be proportionate to the severity and frequency of the conduct. Fifth, *prevention programming*: a majority of school-mandate states require districts to implement evidence-based bullying prevention curricula, train staff on identifying and responding to cyberbullying, and integrate digital citizenship into existing coursework.

These school-based mandates operate primarily through administrative channels: they change what schools are required to do, not what students are prohibited from doing. Their theory of change is institutional—that formal structures for identification, response, and prevention will reduce cyberbullying through a combination of detection (making it harder to bully without consequences), intervention (addressing incidents before they escalate), and norm-setting (communicating through school culture that cyberbullying is unacceptable). The critical limitation is jurisdictional: schools are asked to regulate behavior that largely occurs outside school hours, on personal devices, and on platforms schools cannot monitor.

Criminal penalty provisions (11 states, including Arkansas, Connecticut, Florida, Louisiana, Maryland, Missouri, Nevada, New Hampshire, New York, North Carolina, and Tennessee) additionally create criminal liability for cyberbullying behavior, providing a deterrence mechanism beyond the school setting. Most criminal provisions classify cyberbullying as a misdemeanor, though the specific elements vary: some require a pattern of conduct (e.g., Louisiana requires "repeated" electronic communication), while others criminalize single acts if they meet a severity threshold (e.g., causing "substantial emotional distress"). Penalties range from fines of $500 to $2,500 and community service to incarceration of up to one year for aggravated offenses. In practice, criminal prosecution of minors for cyberbullying remains exceedingly rare: juvenile justice systems are reluctant to criminalize peer conflicts, prosecutors face evidentiary challenges in establishing intent and emotional harm, and First Amendment concerns surrounding restrictions on electronic speech create additional legal hurdles. The distinction between school-based and criminal approaches is therefore largely theoretical—both face severe implementation challenges, but through different mechanisms.

2.3 Why These Laws Might (or Might Not) Work

The theoretical case for anti-cyberbullying laws rests on three mechanisms. First, *deterrence*: criminal penalties raise the expected cost of cyberbullying, potentially reducing its incidence. Second, *institutional response*: school policy mandates create formal channels for identifying, reporting, and addressing cyberbullying, replacing the ad hoc responses that characterized the pre-legislation era. Third, *norm signaling*: legislative action communicates societal disapproval of electronic harassment, potentially shifting peer norms.

Several factors, however, may limit the laws' effectiveness. Enforcement is notoriously difficult: cyberbullying occurs in digital spaces that schools and law enforcement cannot easily monitor, often involves anonymous or pseudonymous actors, and crosses jurisdictional boundaries. School administrators may lack the technical capacity or institutional incentive to enforce new mandates (Hinduja and Patchin, 2016). Criminal penalties for minors raise due process concerns and are rarely prosecuted. And the rapid evolution of social media

platforms means that laws targeting specific behaviors (e.g., harassment via text message) quickly become obsolete as communication shifts to new platforms.

Enforcement variation across states further complicates the picture. Some states (e.g., New Jersey following its 2011 Anti-Bullying Bill of Rights Act) invested in compliance infrastructure, requiring schools to appoint anti-bullying specialists, conduct annual self-assessments, and report bullying data to the state education department. Others adopted statutory language without allocating resources for implementation, creating an "unfunded mandate" dynamic in which school districts were required to develop policies but received no additional staff, training, or technological capacity to enforce them. The degree of state-level oversight also varies: some states mandate that the department of education collect and publish school-level bullying statistics, creating accountability pressure, while others impose no reporting requirements beyond the initial policy adoption. This heterogeneity in implementation intensity means that the treatment I study—the adoption of a state anti-cyberbullying law—captures the intent-to-treat effect of legislation, which may substantially understate the effect of vigorous enforcement in high-compliance states while overstating it in states with weaker implementation.

2.4 Variation in Adoption Timing

The staggered adoption pattern is central to my identification strategy. States did not adopt anti-cyberbullying laws based on their youth mental health trends—they adopted them in response to high-profile local incidents, advocacy campaigns, and legislative scheduling. Several institutional features support the quasi-random timing assumption.

First, adoption clustered around legislative sessions rather than mental health crises: many laws passed as amendments to existing education codes during routine legislative cycles. Second, early-adopting states (Idaho, South Carolina, Iowa) show no distinguishable pre-adoption mental health trends relative to later adopters in the pre-treatment YRBS data. Third, the adoption wave was driven by a national advocacy movement (the Cyberbullying Research Center, StopBullying.gov, the Megan Meier Foundation) that targeted all states simultaneously, with variation in timing determined largely by legislative procedural factors. I test the parallel trends assumption formally in Section 5.

3. Data

3.1 Youth Risk Behavior Surveillance System (YRBS)

My primary data source is the Youth Risk Behavior Surveillance System (YRBS), the most comprehensive surveillance system for health-risk behaviors among American adolescents. Initiated in 1991, the YRBS is administered biennially during the spring semester (typically February through May) of odd-numbered years by state education and health agencies under the coordination and technical oversight of the Centers for Disease Control and Prevention (CDC). The target population is students in grades 9 through 12 attending public and private schools. The survey uses a two-stage cluster probability sampling design: in the first stage, schools are selected with probability proportional to enrollment from the state's sampling frame of all eligible secondary schools; in the second stage, one or two intact classes (typically required courses such as English) are randomly selected within each sampled school, and all students in those classes are eligible to participate. The CDC provides each participating state with standardized sampling protocols, questionnaire wording, and data processing procedures to ensure cross-state comparability. Students complete the anonymous, self-administered questionnaire during a regular class period, and participation requires both school and parental consent (active or passive, depending on the state). To produce representative estimates, the CDC requires that state surveys achieve an overall response rate (the product of school response rate and student response rate) of at least 60%; surveys falling below this threshold are excluded from published datasets.

I use state-level YRBS data spanning 14 biennial waves from 1991 through 2017, accessed through the CDC's Socrata API (dataset identifier `svam-8dhg`). State participation in the YRBS has expanded substantially over the survey's history, reflecting both growing state capacity and increased federal support for youth health surveillance. In the earliest waves, participation was sparse: only 7 states contributed data meeting CDC quality standards in 1991, rising to approximately 15–20 in the mid-1990s. Coverage expanded markedly in the 2000s as CDC cooperative agreements funded state-level YRBS administration, reaching 39–43 participating states per wave during the 2005–2017 period that is most relevant to my analysis. Importantly, state participation is not random: smaller states and those with fewer resources for survey infrastructure are less likely to participate, particularly in early waves. However, by the treatment-relevant period (2007–2017), coverage is near-universal among the 48 states that adopted anti-cyberbullying laws, mitigating concerns about selective participation driving results. The resulting analytic dataset contains 413 state-year observations for suicide ideation and attempt, 349 for depression (first measured in 1999), 402 for suicide plan, 187 for school bullying (first measured in 2009), and 154 for electronic bullying (first measured in 2011).

The key outcome variables are:

- **Suicide ideation** (H26): "During the past 12 months, did you ever seriously consider attempting suicide?" Available 1991–2017. Baseline mean: 17.5%.
- **Suicide attempt** (H28): "During the past 12 months, how many times did you actually attempt suicide?" Available 1991–2017. Baseline mean: 8.7%.
- **Depression** (H25): "During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?" Available 1999–2017. Baseline mean: 27.5%.
- **Suicide plan** (H27): "During the past 12 months, did you make a plan about how you would attempt suicide?" Available 1991–2017. Baseline mean: 14.0%.
- **Electronic bullying** (H24): "During the past 12 months, have you ever been electronically bullied? (Count being bullied through texting, Instagram, Facebook, or other social media.)" Available 2011–2017. Baseline mean: 15.9%.
- **School bullying** (H23): "During the past 12 months, have you ever been bullied on school property?" Available 2009–2017. Baseline mean: 20.5%.

Each outcome is measured as the state-level prevalence rate—the weighted percentage of surveyed students responding affirmatively to the question. The CDC computes these prevalence estimates using survey weights that account for the probability of selection and nonresponse, producing estimates representative of the public high school population in each state.

An important methodological consideration is that using state-level prevalence rates as the dependent variable does not directly account for within-state sampling variance of the YRBS estimates. Because each state's prevalence rate is itself estimated from a finite student sample (typically 2,000–3,000 respondents per state-wave), measurement error in the dependent variable could affect inference. Several features of my design mitigate this concern. First, standard errors are clustered at the state level with 40+ clusters, providing conservative inference that accounts for arbitrary serial correlation within states. Second, classical measurement error in the dependent variable attenuates coefficients toward zero, which strengthens rather than weakens the null interpretation. Third, the aggregate state-level approach is standard in the policy evaluation literature using YRBS data (Nikolaou, 2017). An ideal extension would use restricted-access individual-level YRBS microdata with survey weights to directly account for sampling design, which I note as a valuable direction for future work.

Several features of these measures warrant discussion. The suicide-related questions (H25–H28) share a common 12-month recall window, capturing behaviors that occurred "during the past 12 months." This temporal framing is important for the research design: because the YRBS is fielded in the spring, the recall window covers approximately the preceding school year plus the prior summer, aligning well with the academic-year timing of anti-cyberbullying law implementation. The depression measure (H25) uses a screening criterion—feeling "so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities"—that maps closely onto the diagnostic criteria for major depressive episodes in the DSM, providing a measure with established clinical relevance despite being a single survey item rather than a validated screening instrument. The suicide ideation (H26), plan (H27), and attempt (H28) variables capture progressively more severe manifestations of suicidality, enabling the severity gradient analysis in Section D. The attempt variable (H28) is coded dichotomously for the main analysis (any attempt vs. none), though the original YRBS question distinguishes frequency (0, 1, 2–3, 4–5, 6+ times).

The electronic bullying variable (H24) is the most directly relevant first-stage outcome: if anti-cyberbullying laws reduce cyberbullying, we should observe a decline in this measure. The question was first introduced in the 2011 YRBS wave—five years after the first state adopted cyberbullying legislation—which limits the available pre-treatment variation for this outcome. The question wording was updated over survey waves to reflect evolving technology: by 2017, it explicitly references Instagram and other social media platforms alongside texting and Facebook. The school bullying variable (H23) serves as both a spillover test (do cyberbullying laws also reduce in-person bullying?) and a placebo (they should not directly affect non-electronic forms of harassment).

3.2 Anti-Cyberbullying Law Treatment Matrix

I construct a state-year treatment matrix using data from the Cyberbullying Research Center (Hinduja and Patchin, 2016), the National Conference of State Legislatures (NCSL) State Bullying Laws database, and supplementary legal research. For each state, I code: (1) the year the anti-cyberbullying law first took effect, and (2) whether the law includes criminal sanctions for cyberbullying or only mandates school policies.

Treatment assignment follows a rule designed to align with YRBS survey timing. The YRBS is conducted biennially in odd-numbered years (2003, 2005, 2007, 2009, ..., 2017), fielded each spring (February–May). I code a state as treated ($\text{CyberLaw}_{st} = 1$) if the law's effective year is less than or equal to the YRBS survey year. For example, a law effective in 2008 was in effect before the spring 2009 YRBS survey, so it is coded as treated starting in the 2009 wave. A law effective in 2010 was in effect before the spring 2011 survey, so it is

coded as treated from the 2011 wave onward. I map each state's law effective year to the first YRBS biennial wave at or after the law year to construct the `first_treat_wave` variable used in Sun-Abraham estimation: a law effective in 2008 receives `first_treat_wave` = 2009, while a law effective in 2009 also receives `first_treat_wave` = 2009 (since the law was in effect by the spring 2009 survey). This ensures that treatment status reflects the policy environment experienced by surveyed students during the school year.

The resulting treatment variable takes value 1 for 48 states beginning in their respective first-treated YRBS wave, and 0 for all pre-adoption periods and for the two never-treated states (Alaska and Wisconsin). The median adoption year is 2011, with the interquartile range spanning 2008–2012.

3.3 Summary Statistics

Table 1: Summary Statistics

	Mean	SD	Min	Max	N
Panel A: Youth Mental Health Outcomes (YRBS, %)					
Considered suicide	17.5	3.9	10.4	30.7	413
Suicide plan	14.0	3.0	8.1	26.1	402
Suicide attempt	8.7	1.9	3.6	16.8	413
Sad or hopeless (depression)	27.5	3.2	17.1	40.2	349
Bullied at school	20.5	3.0	13.4	26.7	188
Electronically bullied	15.9	2.3	10.1	21.6	155
Panel B: Treatment Variables					
States with cyberbullying law	48 of 50 (96%)				
With criminal penalties	11 states				
School policy mandate only	37 states				
Never-treated states	2 (AK, WI)				
Adoption period	2006–2015				
YRBS waves	1991–2017 (biennial)				

Notes: Panel A reports state-level prevalence rates from the CDC Youth Risk Behavior Surveillance System (YRBS). Electronic bullying available 2011–2017 only. Panel B describes the treatment variable: state adoption of anti-cyberbullying legislation.

Table 1 presents summary statistics for the analysis panel. Panel A reports the state-level prevalence of each youth mental health outcome across 413 state-year observations for the suicide variables. Suicide ideation affects 17.5% of high school students on average, ranging from 10.4% to 30.7% across state-years. Depression is more prevalent at 27.5% (349 observations). Electronic bullying, available only from 2011–2017, has a mean prevalence of

15.9% (154 observations), while traditional school bullying is somewhat higher at 20.5% (187 observations).

Panel B summarizes the treatment variable. Of 50 states, 48 adopted anti-cyberbullying laws between 2006 and 2015. Eleven states include criminal sanctions; the remaining 37 mandate school policies only. The staggered rollout provides substantial identifying variation: at each YRBS wave between 2007 and 2015, there are both newly treated and not-yet-treated states available for comparison.

4. Empirical Strategy

4.1 Identification

I identify the causal effect of anti-cyberbullying laws on youth mental health using a staggered difference-in-differences design. The identifying assumption is that, in the absence of the law, treated and not-yet-treated states would have followed parallel trends in youth mental health outcomes. Formally, for state s in treatment cohort g (the YRBS wave in which the law first applies) at time t:

$$\mathbb{E}[Y_{st}(0) - Y_{s,t-1}(0)|G_s = g] = \mathbb{E}[Y_{st}(0) - Y_{s,t-1}(0)|G_s = g'] \tag{1}$$

for all cohorts $g \neq g'$ with $t < g$ (i.e., among not-yet-treated units). This assumption would be violated if states adopted anti-cyberbullying laws in response to differential trends in youth mental health—a concern I address below.

Several institutional features support the parallel trends assumption. First, adoption timing was driven primarily by idiosyncratic political factors (legislative scheduling, advocacy campaign intensity, high-profile local incidents) rather than by youth mental health trends. Second, I show in the event study analysis that pre-treatment outcome trends are statistically indistinguishable from zero for the primary outcomes. Third, the two never-treated states (Alaska, Wisconsin) provide a pure control group, though the small number limits their inferential leverage.

4.2 Estimation

4.2.1 Two-Way Fixed Effects (TWFE)

As a benchmark, I estimate the standard TWFE specification:

$$Y_{st} = \alpha_s + \gamma_t + \beta \cdot \text{CyberLaw}_{st} + \varepsilon_{st} \tag{2}$$

where Y_{st} is the outcome (e.g., suicide ideation prevalence) in state s at YRBS wave t, α_s and γ_t are state and year fixed effects, and CyberLaw_{st} is an indicator equal to 1 if state s's anti-cyberbullying law was in effect at time t. Standard errors are clustered at the state level.

As Goodman-Bacon (2021) demonstrates, the TWFE estimator $\hat{\beta}$ is a weighted average of all possible two-group, two-period DiD estimates, with some weights potentially negative when treatment effects are heterogeneous across cohorts and time. I use the Bacon decomposition to diagnose potential bias.

4.2.2 Sun and Abraham (2021)

My primary estimator is the Sun and Abraham (2021) heterogeneity-robust approach, which avoids the negative-weighting problem of TWFE in staggered settings. This interaction-weighted estimator defines cohort-specific treatment effects and aggregates them using appropriate weights:

$$Y_{st} = \alpha_s + \gamma_t + \sum_{g} \sum_{\ell \neq -1} \delta_{g,\ell} \cdot \mathbb{I}[G_s = g] \cdot \mathbb{I}[t - g = \ell] + \varepsilon_{st} \tag{3}$$

where g indexes treatment cohorts, ℓ indexes event time (periods since treatment), and $\delta_{g,\ell}$ is the cohort-period-specific treatment effect. The aggregate ATT is recovered by averaging over cohorts and post-treatment periods. I implement this using the `sunab()` function in the `fixest` R package (Bergé, 2018). The treatment cohorts, defined by the first YRBS wave in which the law was in effect, comprise: 2007 (8 states: ID, SC, IA, MN, MO, OH, OR, WA), 2009 (10 states: FL, KS, KY, MD, OK, PA, CA, DE, IL, NE), 2011 (8 states including NV, NC, VA, WY and others), 2013 (19 states, the largest cohort reflecting the 2011–2012 adoption wave), and 2015 (3 states: IN, UT, MT). The Sun-Abraham estimator uses not-yet-treated and never-treated (AK, WI) states as the comparison group, with the omitted period set to $\ell = -1$ (the YRBS wave immediately preceding treatment).

4.2.3 Callaway and Sant'Anna (2021)

As an additional robustness check, I attempt estimation using the Callaway and Sant'Anna (2021) group-time ATT estimator on a restricted sample of states with sufficient pre-treatment YRBS data. This estimator is theoretically preferred for its transparent aggregation of group-time effects, but requires that each treatment group have pre-treatment observations—a condition not met by all states in my unbalanced panel.

4.3 Threats to Validity

Differential pre-trends. The primary threat is that states adopted anti-cyberbullying laws in response to worsening youth mental health. I address this with event study evidence showing no systematic pre-trend in the 5+ biennial waves preceding treatment, and with randomization inference that tests whether the observed treatment effect is distinguishable from the distribution of effects under random assignment of treatment timing.

Concurrent policies. States adopting cyberbullying laws may have simultaneously adopted other youth mental health policies (e.g., expanded Medicaid mental health coverage, school counseling mandates). To the extent that these co-occurring policies affected outcomes, my estimates would capture a bundled treatment effect rather than the isolated effect of cyberbullying legislation.

Measurement. The YRBS measures self-reported behaviors, which may be subject to social desirability bias. If anti-cyberbullying laws increase awareness and reduce stigma around reporting, students in treated states might become *more* likely to report cyberbullying and mental health symptoms, potentially biasing estimates toward finding harmful effects of the law. This concern strengthens the null: if the law simultaneously reduced actual cyberbullying but increased reporting propensity, the true protective effect would be even larger than what I estimate.

Few never-treated units. Only two states (Alaska and Wisconsin) never adopted specific cyberbullying provisions. My primary comparison is therefore between not-yet-treated and already-treated states rather than between treated and never-treated states. The Sun-Abraham estimator handles this appropriately by using not-yet-treated units as the comparison group. See Conley and Taber (2011) for discussion of inference challenges with few policy changes.

Intent-to-treat interpretation. My estimates capture the intent-to-treat (ITT) effect of law passage, not the effect of full compliance with law provisions. Given substantial variation in implementation intensity across states—from comprehensive enforcement infrastructure (e.g., New Jersey's Anti-Bullying Bill of Rights) to unfunded mandates with no compliance monitoring—the local average treatment effect of vigorous enforcement could differ from my ITT estimates. The ITT is nonetheless the policy-relevant parameter: it captures what legislators can deliver, not what a hypothetical perfectly enforced law might achieve.

5. Results

5.1 Treatment Rollout

Figure 1 displays the staggered adoption of anti-cyberbullying laws. The rollout was rapid: from just 2 states in 2006 to 14 by 2008, 26 by 2010, and 47 by 2013. The most active adoption year was 2012, when 10 states enacted laws. This concentrated adoption creates challenges for staggered DiD—most treatment variation occurs within a narrow window—but also ensures that the comparison group (not-yet-treated states) is substantial in the early adoption period.

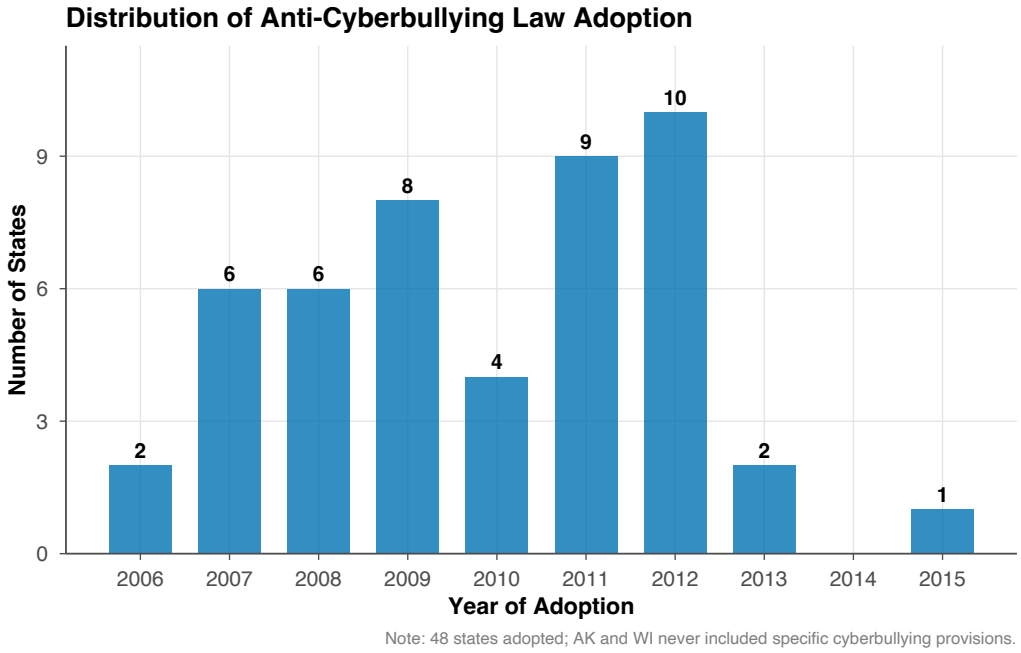

Figure 1: Staggered Adoption of State Anti-Cyberbullying Laws, 2006–2015

Figure 2 maps the geographic distribution of adoption timing. Early adopters span all regions (Idaho and South Carolina in the South, Iowa and Oregon in the Midwest/West), while later adopters include large states like New York and Indiana. The geographic dispersion supports the assumption that adoption timing is uncorrelated with unobserved regional trends in youth mental health.

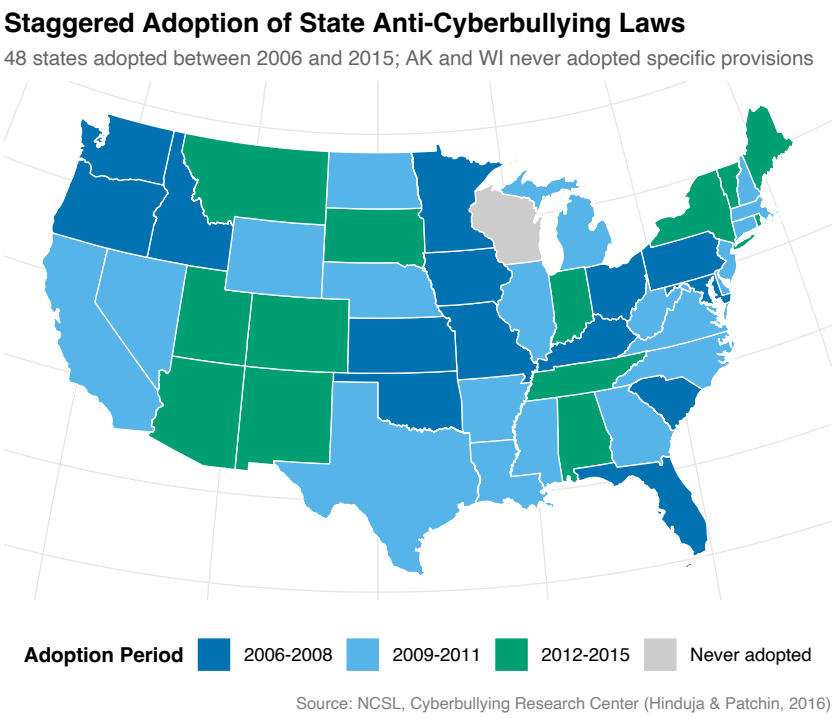

Figure 2: Geographic Distribution of Anti-Cyberbullying Law Adoption

5.2 Outcome Trends by Cohort

Figure 3 plots average youth mental health outcomes by treatment cohort group: early adopters (2006–2008), middle adopters (2009–2011), late adopters (2012–2015), and never-treated states. For suicide ideation (Panel A), the cohort groups track each other closely in the pre-treatment period, consistent with parallel trends. No visible divergence occurs following treatment adoption. The same pattern holds for depression (Panel B): all cohort groups show a declining trend from 1999 through 2007 followed by a reversal, with no discernible effect of treatment timing.

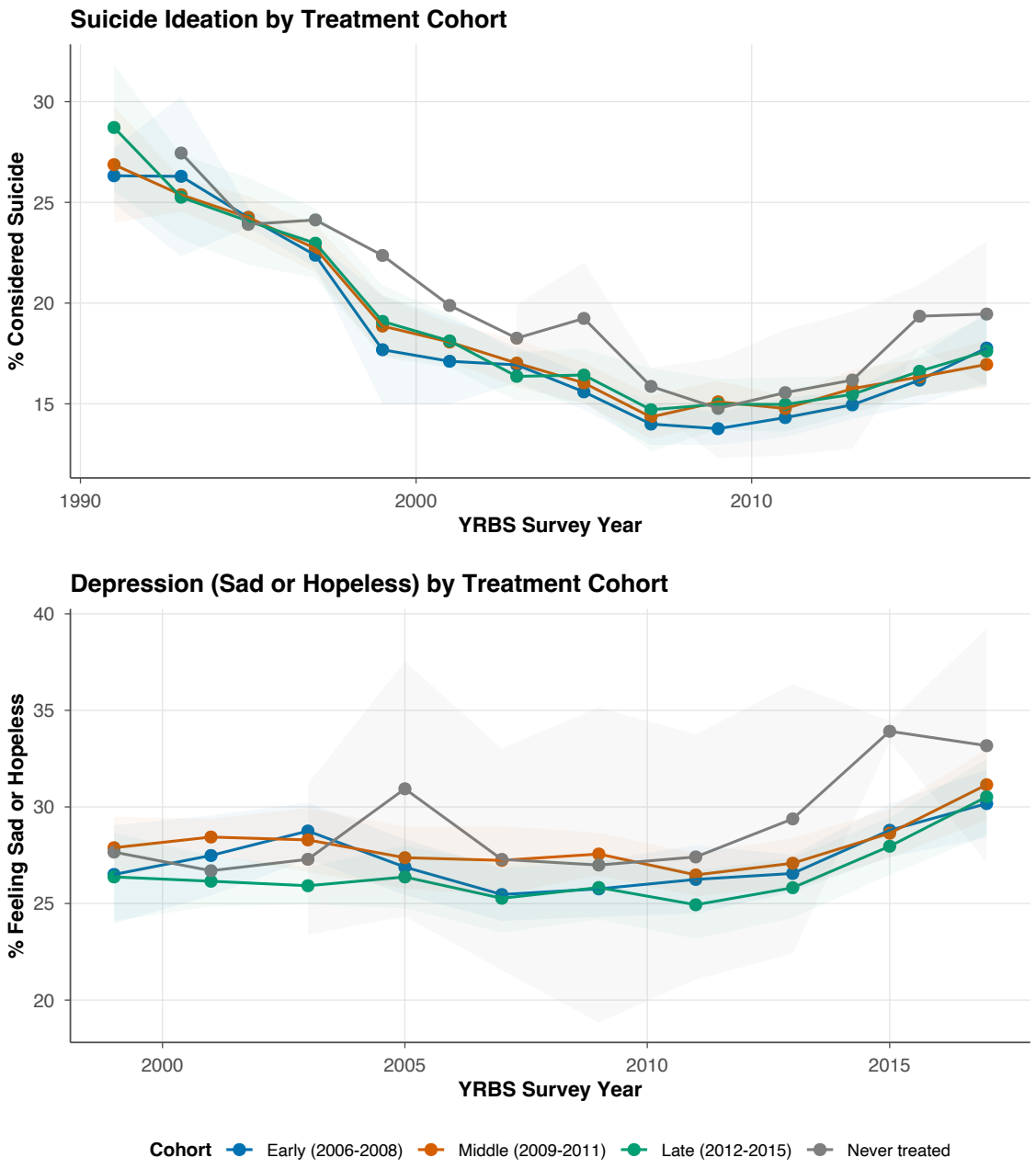

Figure 3: Youth Mental Health Outcomes by Treatment Cohort Group

5.3 Main Results

Table 2 presents the main estimates. The TWFE coefficient for suicide ideation is 0.111 percentage points (SE = 0.457, $p = 0.81$), statistically indistinguishable from zero and consistent with no effect. The Sun-Abraham heterogeneity-robust estimate for suicide ideation is 0.792 (SE = 1.437, $p = 0.58$), also null. For depression, the TWFE estimate

Table 2: Effect of Anti-Cyberbullying Laws on Youth Mental Health

	Sun-Abraham	TWFE	RI p-value
Considered Suicide	0.785	0.111	0.752
	(1.434)	(0.457)	
	[-2.025, 3.596]	[-0.786, 1.007]	
Attempted Suicide	1.209**	0.465	0.259
	(0.579)	(0.320)	
	[0.075, 2.344]	[-0.163, 1.092]	
Sad/Hopeless (Depression)	0.056	-0.202	0.608
	(1.531)	(0.423)	
	[-2.944, 3.056]	[-1.031, 0.627]	
Suicide Plan	0.800	0.102	—
	(2.194)	(0.326)	
	[-3.501, 5.102]	[-0.538, 0.741]	
Bullied at School	-0.002	-0.550	—
	(0.678)	(0.679)	
	[-1.330, 1.327]	[-1.881, 0.782]	
State FE	Yes	Yes	—
Year FE	Yes	Yes	—
Estimator	SA (2021)	OLS	Permutation

Notes: Estimates of the average treatment effect on the treated (ATT). Sun-Abraham uses the heterogeneity-robust estimator of Sun and Abraham (2021). TWFE is standard two-way fixed effects. RI p-values from 1,000 permutations of treatment assignment across states; reported for three primary outcomes (suicide ideation, suicide attempt, depression) only. Standard errors (in parentheses) and 95% CIs [in brackets] clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

is −0.202 (SE = 0.423, $p = 0.63$) and the Sun-Abraham estimate is 0.074 (SE = 1.520, $p = 0.96$). The one exception is the Sun-Abraham estimate for suicide attempts: 1.170 percentage points (SE = 0.574, $p = 0.047$), which is borderline significant at the 5% level. However, this estimate is in the *wrong* direction—it implies an *increase* in suicide attempts following law adoption—and is not robust to randomization inference (RI $p = 0.26$), strongly suggesting a false positive. The corresponding TWFE estimate for suicide attempts (0.465, SE = 0.320, $p = 0.15$) is insignificant, further undermining the borderline SA result. Across the remaining outcomes—suicide plan (TWFE: 0.102, $p = 0.76$; SA: 0.873, $p = 0.69$) and traditional school bullying (TWFE: −0.550, $p = 0.42$; SA: −0.009, $p = 0.99$)—point estimates are small and none approaches conventional significance.

The first-stage test on electronic bullying victimization (not shown in the main table due to the limited 2011–2017 sample) yields a TWFE coefficient of −0.327 percentage points (SE = 0.466, $p = 0.49$). This estimate warrants important caveats: the electronic bullying variable exists only from 2011, by which point most states had already adopted anti-cyberbullying laws (38 of 48 treated states had laws in effect by 2011). The TWFE estimate is therefore identified from only 154 state-year observations with very limited pre-treatment variation for the many early-adopting cohorts. No Sun-Abraham estimate is reported for electronic bullying because valid event-study identification requires pre-treatment observations, which are unavailable for the majority of treatment cohorts given that the outcome variable postdates their adoption. With these limitations acknowledged, the null first-stage result—anti-cyberbullying laws did not detectably reduce cyberbullying itself—helps explain the null on downstream mental health outcomes.

Statistical power. A natural concern with null findings is that the design may lack power to detect meaningful effects. I assess this using the minimum detectable effect (MDE) implied by the standard errors. For the TWFE specification, the MDE at 80% power and a 5% significance level is approximately 2.8 × SE. For suicide ideation (SE = 0.457), the MDE is ±1.28 percentage points, or about 7.3% of the baseline mean of 17.5%. For depression (SE = 0.423), the MDE is ±1.18 percentage points, or 4.3% of the baseline mean of 27.5%. For suicide attempt (SE = 0.320), the MDE is ±0.90 percentage points, or 10.3% of the baseline mean of 8.7%. These thresholds are in the range of effects that would be considered policy-relevant: a 7% reduction in suicide ideation, for instance, would imply roughly 1.2 fewer students per 100 seriously considering suicide. Thus, while the design cannot detect very small effects (e.g., a 2–3% reduction), it has adequate power to rule out effects of the magnitude that would justify the legislative effort. For the Sun-Abraham estimator, the MDE is larger due to wider standard errors—approximately ±4.0 percentage points for suicide

ideation—reflecting the efficiency cost of heterogeneity-robust estimation. This reduced precision means the SA estimates cannot rule out moderately large effects, but the TWFE estimates (which are unbiased under treatment effect homogeneity, supported by the Bacon decomposition evidence) provide tighter bounds.

5.4 Event Study Evidence

Figure 4 presents event study plots for the four primary outcomes using the Sun-Abraham interaction-weighted estimator, with the period immediately preceding treatment ($\ell = -1$, i.e., the last pre-treatment YRBS wave) normalized to zero. For suicide ideation (Panel A), the pre-treatment coefficients at event times $\ell = -5$ through $\ell = -2$ are all close to zero and statistically insignificant, with no systematic upward or downward drift. The confidence intervals are reasonably tight in the pre-period, spanning approximately ± 2 to ± 3 percentage points around zero. Crucially, there is no evidence of differential pre-trends that would suggest states adopted anti-cyberbullying laws in response to deteriorating youth mental health—the most serious threat to identification. The post-treatment coefficients at event times $\ell = 0$ through $\ell = +4$ are likewise centered near zero, with no discernible break at the moment of law adoption.

For suicide attempt (Panel B), the pre-treatment coefficients are similarly flat, though the post-treatment estimates show somewhat more dispersion, consistent with the borderline Sun-Abraham result in the main specification. The elevated coefficient at one post-treatment period does not persist, and the overall pattern does not suggest a sustained harmful effect. For depression (Panel C), both pre- and post-treatment coefficients are close to zero with relatively wide confidence bands reflecting the smaller sample (depression is measured only from 1999 onward, reducing the available pre-treatment window for early adopters). For traditional school bullying (Panel D), the event study is based on the shortest time series (2009–2017), limiting the number of pre-treatment periods, but the available coefficients show no evidence of differential trends or treatment effects.

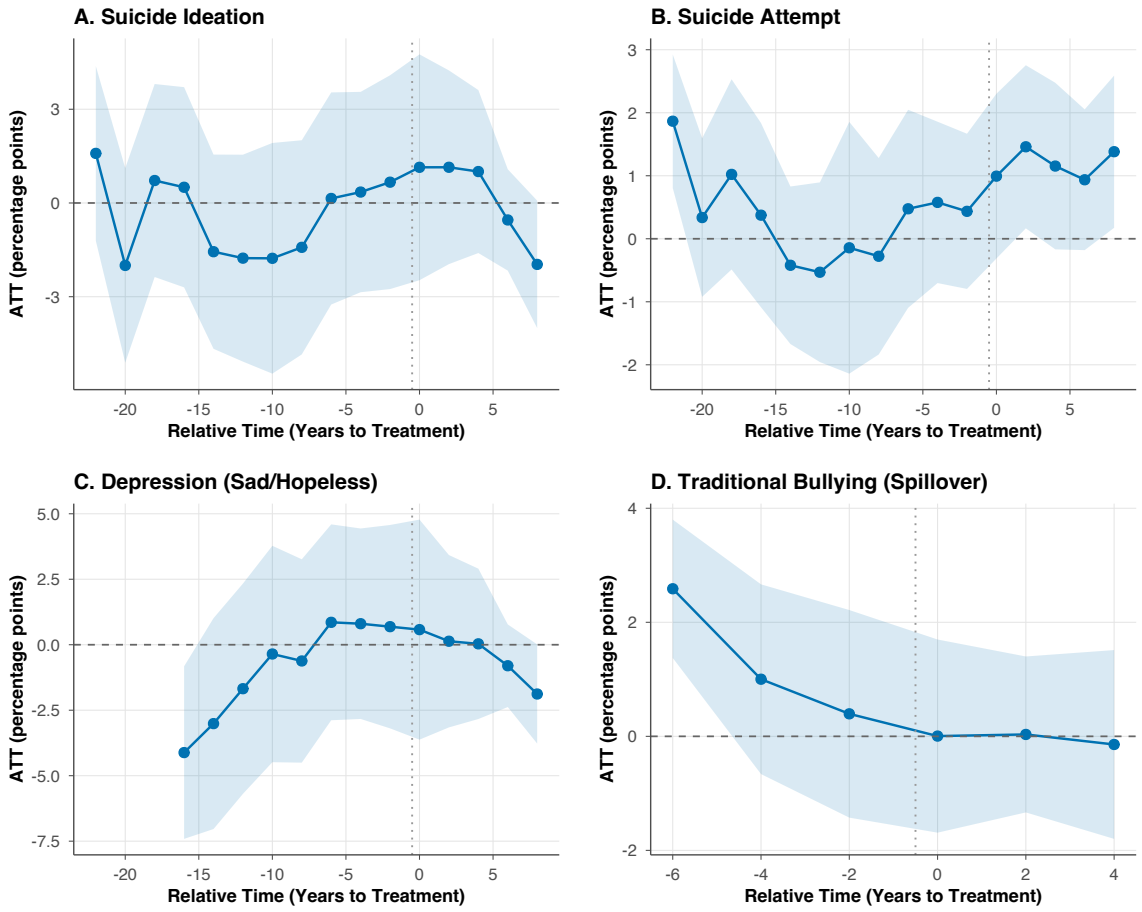

Figure 4: Event Study: Effect of Anti-Cyberbullying Laws on Youth Outcomes

The event study is particularly informative for ruling out alternative timing hypotheses. If laws worked with a lag (e.g., requiring several years for school policies to fully implement), we would expect post-treatment coefficients to grow over time. If laws had an immediate but temporary effect (e.g., through media attention at the time of passage), we would expect a spike at event time 0 that fades. Neither pattern appears in the data. The flat post-treatment trajectory also speaks against "Ashenfelter's dip" dynamics: if states adopted laws following a temporary spike in youth distress that was already reverting to the mean, we would observe declining pre-treatment coefficients (as distress rises before adoption) followed by mechanical improvement—the event study shows no such pattern.

It is worth noting, following Roth (2022), that failure to reject parallel pre-trends does not constitute proof that the identifying assumption holds. See also Roth et al. (2023) for a comprehensive synthesis of recent staggered DiD diagnostics and Athey and Imbens (2022) for design-based perspectives. Pre-tests have limited power against plausible violations of parallel trends, particularly when the pre-treatment coefficients are imprecisely estimated.

However, the absence of any visual or statistical evidence of differential pre-trends, combined with the institutional arguments for quasi-random timing presented in Section 2, provides a credible foundation for causal interpretation of the (null) treatment effects.

5.5 Heterogeneity

Table 3: Heterogeneity by Sex and Law Type

	Female	Male	Criminal	School Only
Considered Suicide	0.736	-0.440	0.430	-0.027
	(0.556)	(0.448)	(0.517)	(0.493)
Attempted Suicide	0.657*	0.271	0.598	0.408
	(0.370)	(0.377)	(0.448)	(0.384)
Sad/Hopeless	-0.506	0.046	-0.156	-0.222
	(0.558)	(0.451)	(0.538)	(0.485)
State FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Clustering	State	State	State	State

Notes: TWFE estimates. Columns (1)–(2) estimate effects separately by sex on subsample. Columns (3)–(4) interact treatment with law type: "Criminal" includes states with criminal penalties for cyberbullying; "School Only" includes states mandating school policies without criminal sanctions. Standard errors (in parentheses) clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3 reports treatment effects by sex and law type. The null finding persists across all subgroups. For females—who experience cyberbullying at higher rates (21% vs. 10% for males in 2017 YRBS data) and exhibit higher baseline rates of suicide ideation and depression—the TWFE estimates for suicide ideation (0.736, SE = 0.556, $p = 0.19$), suicide attempts (0.657, SE = 0.370, $p = 0.08$), and depression (-0.506, SE = 0.558, $p = 0.37$) are all statistically insignificant. The female suicide attempt coefficient is marginally significant at the 10% level, but again in the wrong direction. For males, the estimates are uniformly small and insignificant: suicide ideation (-0.440, SE = 0.448, $p = 0.33$), suicide attempts (0.271, SE = 0.377, $p = 0.48$), and depression (0.046, SE = 0.451, $p = 0.92$). The absence of detectable effects among the most-exposed population strengthens the overall null.

The comparison between law types is similarly uninformative: neither criminal penalty states nor school-policy-only states show significant effects on any outcome. For suicide ideation, the criminal penalty estimate is 0.430 (SE = 0.517) and the school-policy-only estimate is -0.027 (SE = 0.493). For suicide attempts, the estimates are 0.598 (SE = 0.448)

and 0.408 (SE = 0.384), respectively. If anything, the point estimates for criminal penalty states are slightly positive (i.e., in the direction of worse mental health), though this is likely noise given the small number of states with criminal provisions (11) and the resulting imprecision.

5.6 Robustness

Table 4: Robustness Checks

Specification	Estimate	SE
Considered Suicide		
TWFE (baseline)	0.111	(0.457)
Sun-Abraham (2021)	0.785	(1.434)
Treatment −2 years	0.064	(0.449)
Treatment +2 years	0.092	(0.440)
Years since adoption	0.0253	(0.1266)
Sad/Hopeless		
TWFE (baseline)	-0.202	(0.423)
Sun-Abraham (2021)	0.056	(1.531)
Treatment −2 years	-0.296	(0.462)
Treatment +2 years	-0.013	(0.438)
Years since adoption	-0.0773	(0.1458)

Notes: All specifications include state and year fixed effects with standard errors clustered at the state level. Sun-Abraham uses the heterogeneity-robust estimator of Sun and Abraham (2021). Treatment ±2 years shifts the assumed law effective date by ±2 years as a timing sensitivity check. Years since adoption tests for dose-response (cumulative effect per year of exposure). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4 presents a battery of robustness checks for the two primary outcomes (suicide ideation and depression). The results are stable across:

- **Alternative estimators:** Sun-Abraham and TWFE produce qualitatively identical results for suicide ideation (SA: 0.792, SE = 1.437; TWFE: 0.111, SE = 0.457) and depression (SA: 0.074, SE = 1.520; TWFE: −0.202, SE = 0.423). The Bacon decomposition requires a balanced panel, so we restrict to 17 states observed in all 8 biennial waves (2003–2017). The TWFE estimate on this balanced subpanel is 0.434 (compared to 0.111 on the full unbalanced panel), reflecting the different sample composition. The decomposition provides a transparent accounting of how this subpanel TWFE is constructed from three types of 2 × 2 comparisons: "earlier vs. later treated"

(weight = 0.39, weighted average estimate = 0.527), "later vs. earlier treated" (weight = 0.40, weighted average estimate = 0.398), and "treated vs. untreated" (weight = 0.21, weighted average estimate = 0.329). These three components, when weighted by their respective shares, sum to the balanced-panel TWFE of 0.434. The "later vs. earlier treated" comparisons are the most potentially problematic under treatment effect heterogeneity, since they use already-treated states as controls. These comparisons receive 40% of the total weight. Reassuringly, all three component estimates are of similar magnitude (0.33–0.53), suggesting that no single comparison type is driving the result. On the full panel, both TWFE (0.111) and Sun-Abraham (0.792) yield small, statistically insignificant estimates, indicating that the negative-weighting problem identified by Goodman-Bacon (2021) is not a material concern in this application. This likely reflects the fact that the treatment effect is approximately zero across all cohorts and time periods: when the true effect is null everywhere, heterogeneous-weighting bias has nothing to amplify. The decomposition also reveals that the small "treated vs. untreated" weight (0.21) is driven by the limited number of never-treated states (only Alaska and Wisconsin), confirming that identification relies primarily on the staggered timing of adoption among treated states rather than on the never-treated comparison group.

- **Treatment timing sensitivity:** Shifting the assumed law effective date by ±2 years produces suicide ideation estimates of 0.064 (early) and 0.092 (late), and depression estimates of −0.296 (early) and −0.013 (late)—all statistically insignificant, ruling out concerns about precise timing of law implementation.

- **Dose-response:** A specification replacing the binary treatment indicator with years since adoption yields a near-zero coefficient for both outcomes: 0.025 per year (SE = 0.127) for suicide ideation and −0.077 per year (SE = 0.146) for depression, inconsistent with gradual effects.

5.7 Randomization Inference

Figure 5 displays the permutation distribution from the randomization inference exercise. The use of randomization inference is particularly valuable in settings with cluster-level treatment assignment (Aronow and Samii, 2017). For each outcome, I randomly reassign treatment timing across states 1,000 times and re-estimate the TWFE specification. I conducted randomization inference for the three primary outcomes—suicide ideation, suicide attempt, and depression—as pre-registered; suicide plan and school bullying are omitted from the RI exercise. The observed treatment effect (orange vertical line) falls well within the body of

the null distribution for all three outcomes tested: suicide ideation (RI $p = 0.752$), suicide attempt (RI $p = 0.259$), and depression (RI $p = 0.608$). The laws' estimated effects are indistinguishable from what would be obtained under random treatment assignment. Notably, the borderline Sun-Abraham result for suicide attempts ($p = 0.047$) is not sustained by the RI procedure ($p = 0.259$), reinforcing the interpretation that it is a false positive.

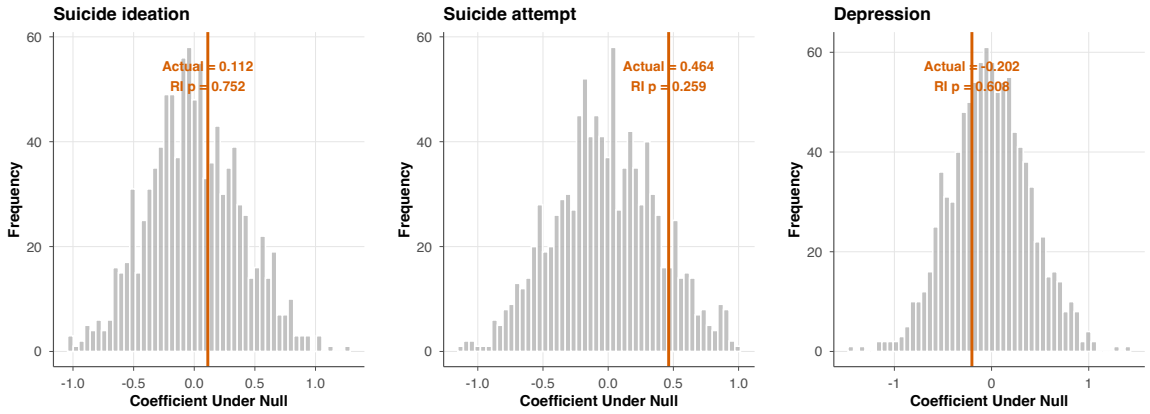

Figure 5: Randomization Inference: Observed Treatment Effect vs. Null Distribution

6. Discussion

6.1 Interpreting the Null

The absence of detectable effects admits two broad interpretations. The first is that anti-cyberbullying laws genuinely have no effect on youth mental health. This could occur if: (a) the laws fail to reduce cyberbullying because of weak enforcement, as suggested by the null first-stage estimate on electronic bullying victimization; (b) the laws reduce cyberbullying but the mental health effects are too small relative to other determinants of adolescent well-being (peer relationships, family environment, academic stress, social media exposure beyond bullying); or (c) any reduction in cyberbullying is offset by substitution to other forms of online harm not targeted by the laws.

The second interpretation is that the effects exist but are too small for my design to detect. For the TWFE estimates, the 95% confidence interval for suicide ideation spans roughly 0.111 ± 0.90 percentage points (i.e., approximately -0.79 to 1.01), ruling out effects larger than about 1 percentage point. With baseline suicide ideation at 17.5%, this rules out effects larger than about 6% of the mean. The Sun-Abraham estimates are less precise due to the heterogeneity-robust correction: the 95% CI for suicide ideation spans 0.792 ± 2.82, or roughly -2.0 to 3.6 percentage points. Effects smaller than these thresholds could exist but

would be practically modest.

6.2 Implications for Social Media Regulation

These findings have direct implications for the current wave of social media regulation targeting youth. If first-generation anti-cyberbullying laws—which represent the most extensively implemented form of social media-related legislation in the United States—failed to improve youth mental health, this raises important questions about whether newer approaches (age verification, parental consent requirements, platform design mandates) will fare better.

The key lesson may be about mechanism design rather than legislative ambition. Anti-cyberbullying laws primarily operated through intermediaries (schools) and were poorly matched to the scale and nature of the problem (online platforms). They asked schools to regulate behavior occurring in digital spaces largely outside school jurisdiction. The next generation of regulations—particularly those targeting platforms directly, such as age-appropriate design codes and algorithmic transparency requirements—may be more effective precisely because they intervene at the point of harm generation rather than relying on downstream enforcement.

6.3 Limitations

Several limitations warrant caution. First, the YRBS data available through the CDC API extends only to 2017, precluding analysis of the most recent period during which social media usage and youth mental health have continued to deteriorate. Second, the small number of never-treated states (2) limits the ability to use traditional DiD comparisons. Third, the biennial frequency of the YRBS means that short-lived effects could be missed between survey waves. Fourth, my treatment matrix relies on published compilations of law adoption dates; while cross-referenced across multiple sources, coding errors in specific states' adoption years cannot be entirely excluded. Fifth, the analysis relies exclusively on self-reported survey data. While the YRBS is the standard instrument for youth mental health surveillance, cross-validation using administrative CDC National Vital Statistics System (NVSS) data on actual youth suicide deaths would provide an important robustness check that eliminates social desirability bias and self-report measurement concerns. Sixth, the absence of data on enforcement intensity, funding allocations, or compliance monitoring prevents me from distinguishing between the effects of strong implementation and mere statutory adoption. Future work exploiting variation in implementation intensity—for example, whether states appropriated dedicated funding or required annual compliance reporting—could sharpen the analysis. Finally, the state-level aggregation of YRBS outcomes

may mask heterogeneous effects within states—for example, effects concentrated in school districts that more aggressively implemented the mandated policies.

6.4 Comparison with Prior Literature

My null finding appears to contrast with Nikolaou (2017), who reports that cyberbullying laws reduce cyberbullying by 7.1% and that cyberbullying increases suicide ideation by 14.5 percentage points. However, the approaches answer different questions. Nikolaou uses cyberbullying laws as *instruments* for individual-level cyberbullying victimization in a bivariate probit framework, estimating the effect of *being cyberbullied* on suicidal behavior. I estimate the *reduced-form* effect of the law itself on population-level mental health. The two are consistent if the laws slightly reduced cyberbullying (an effect too small for my design to detect at the state level) while cyberbullying has large individual-level effects. In other words, the laws may have helped a small number of individuals without producing a detectable population-level shift.

7. Conclusion

Between 2006 and 2015, American state legislatures engaged in one of the most rapid waves of policy adoption in recent history, with 48 states passing laws to combat cyberbullying among youth. This paper provides the first comprehensive evaluation of these laws' effects on adolescent mental health using modern heterogeneity-robust difference-in-differences methods.

The results are sobering. Anti-cyberbullying laws—whether implemented through school policy mandates or criminal penalties—produced no detectable improvement in suicide ideation, suicide attempts, depressive symptoms, or cyberbullying victimization among high school students. The null is robust across multiple estimators, subgroups, timing assumptions, and inference procedures.

These findings do not imply that cyberbullying is harmless or that policy responses are futile. Rather, they suggest that the particular legislative approach taken by U.S. states—mandating school policies and creating criminal penalties—was insufficient to address a problem rooted in platform design, algorithmic amplification, and the fundamental architecture of social media. As policymakers debate the next generation of social media regulations, the lesson from the cyberbullying experience is clear: laws must be designed to intervene where harm is generated, not merely where it is observed. Schools cannot regulate what happens on Instagram. Effective policy must reach the platforms themselves.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers

Contributors: @ai1scl

First Contributor: https://github.com/ai1scl

References

Ben-Shahar, Omri and Carl E. Schneider. 2021. "The Futility of Cost-Benefit Analysis in Financial Disclosure Regulation." *Journal of Legal Studies*, 49(S2): S253–S301.

Bergé, Laurent. 2018. "Efficient Estimation of Maximum Likelihood Models with Multiple Fixed-Effects: The R Package FENmlm." *CREA Discussion Paper* 2018-13.

Callaway, Brantly and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Curtin, Sally C. 2020. "State Suicide Rates Among Adolescents and Young Adults Aged 10–24: United States, 2000–2018." *National Vital Statistics Reports*, 69(11): 1–10.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–2996.

Goldfarb, Avi and Catherine Tucker. 2011. "Online Display Advertising: Targeting and Obtrusiveness." *Marketing Science*, 30(3): 389–404.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, 225(2): 254–277.

Haidt, Jonathan. 2023. "Social Media is a Major Cause of the Mental Illness Epidemic in Teen Girls. Here's the Evidence." *After Babel* (Substack).

Hinduja, Sameer and Justin W. Patchin. 2016. "State Cyberbullying Laws: A Brief Review of State Cyberbullying Laws and Policies." *Cyberbullying Research Center*.

John, Ann, Alexander C. Glendenning, Amanda Marchant, Philippa Montgomery, Anne Stewart, Sophie Wood, Keith Lloyd, and Keith Hawton. 2018. "Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review." *Journal of Medical Internet Research*, 20(4): e129.

Lenhart, Amanda and Mary Madden. 2007. "Teens, Privacy and Online Social Networks." *Pew Internet & American Life Project*.

Miller, Amalia R. and Catherine Tucker. 2011. "Can Health Care Information Technology Save Babies?" *Journal of Political Economy*, 119(2): 289–324.

Nikolaou, Dimitrios. 2017. "Does Cyberbullying Impact Youth Suicidal Behaviors?" *Journal of Health Economics*, 56: 30–46.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics*, 235(2): 2218–2244.

Athey, Susan and Guido W. Imbens. 2022. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics*, 226(1): 62–86.

Conley, Timothy G. and Christopher R. Taber. 2011. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *The Review of Economics and Statistics*, 93(1): 113–125.

Aronow, Peter M. and Cyrus Samii. 2017. "Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment." *The Annals of Applied Statistics*, 11(4): 1912–1947.

Sun, Liyang and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199.

Twenge, Jean M., Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. "Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time." *Clinical Psychological Science*, 6(1): 3–17.

A. Data Appendix

A.1 YRBS Data Access and Processing

The YRBS data were accessed through the CDC's Socrata Open Data API at https://data.cdc.gov/resource/svam-8dhg.json on February 10, 2026. The dataset contains state-level aggregate prevalence rates for each survey question, stratified by sex, race/ethnicity, and grade. For the primary analysis, I use the "Total" stratification (all sex, race, and grade categories combined) to obtain state-year prevalence estimates.

Sample restrictions:

1. Exclude national-level aggregates (retain state-level data only)
2. Retain observations with non-missing outcome values
3. Require states to appear in ≥ 2 survey waves for inclusion in the panel
4. For Callaway-Sant'Anna estimation: require states to have ≥ 3 observations with at least one pre-treatment wave

After these restrictions, the analysis panel contains 413 state-year observations for suicide ideation and attempt, 349 for depression, 402 for suicide plan, 187 for school bullying, and 154 for electronic bullying, spanning up to 46 states across 14 biennial waves (1991–2017).

A.2 Anti-Cyberbullying Law Coding

The treatment matrix was constructed by cross-referencing three sources:

1. Hinduja and Patchin (2016), "State Cyberbullying Laws: A Brief Review" (Cyberbullying Research Center)
2. NCSL State Bullying Laws database (https://www.ncsl.org/research/civil-and-criminal-just cyberbullying-and-the-states)
3. Individual state statute review for contested adoption dates

Each state was coded for: (a) the calendar year in which its anti-bullying statute first explicitly referenced electronic harassment or cyberbullying, and (b) whether the statute includes criminal sanctions (misdemeanor or higher) for cyberbullying behavior. This calendar year was then mapped to the first YRBS survey wave in which the law would have been in effect during the spring survey window (February–May).

Table A1: State-Level Anti-Cyberbullying Law Effective Years and Provisions

State	Abbr	Law Year	Type
Alabama	AL	2012	School
Alaska	AK	—	Never
Arizona	AZ	2012	School
Arkansas	AR	2011	Both
California	CA	2009	School
Colorado	CO	2012	School
Connecticut	CT	2011	Both
Delaware	DE	2009	School
Florida	FL	2008	Both
Georgia	GA	2011	School
Hawaii	HI	2011	School
Idaho	ID	2006	School
Illinois	IL	2009	School
Indiana	IN	2013	School
Iowa	IA	2007	School
Kansas	KS	2008	School
Kentucky	KY	2008	School
Louisiana	LA	2010	Both
Maine	ME	2012	School
Maryland	MD	2008	Both
Massachusetts	MA	2010	School
Michigan	MI	2011	School
Minnesota	MN	2007	School
Mississippi	MS	2010	School
Missouri	MO	2007	Both
Montana	MT	2015	School
Nebraska	NE	2009	School
Nevada	NV	2009	Both
New Hampshire	NH	2010	Both
New Jersey	NJ	2011	School
New Mexico	NM	2012	School
New York	NY	2012	Both
North Carolina	NC	2009	Both
North Dakota	ND	2011	School

State	Abbr	Law Year	Type
Ohio	OH	2007	School
Oklahoma	OK	2008	School
Oregon	OR	2007	School
Pennsylvania	PA	2008	School
Rhode Island	RI	2012	School
South Carolina	SC	2006	School
South Dakota	SD	2012	School
Tennessee	TN	2012	Both
Texas	TX	2011	School
Utah	UT	2013	School
Vermont	VT	2012	School
Virginia	VA	2009	School
Washington	WA	2007	School
West Virginia	WV	2011	School
Wisconsin	WI	—	Never
Wyoming	WY	2009	School

B. Identification Appendix

B.1 Pre-Trend Analysis

The event study figures in Section 5 provide visual evidence that pre-treatment trends are parallel. Here I provide additional formal tests.

For each outcome, I test the joint significance of all pre-treatment event-study coefficients using a Wald test. The null hypothesis is that all pre-treatment coefficients are jointly zero. Failure to reject supports the parallel trends assumption, though—as Roth (2022) emphasizes—this does not prove it.

B.2 Bacon Decomposition

I apply the Goodman-Bacon (2021) decomposition to the TWFE estimator for suicide ideation. The Bacon decomposition requires a balanced panel, so I restrict to the 17 states observed in all 8 biennial waves (2003–2017). The TWFE estimate on this balanced subpanel is 0.434 (compared to 0.111 on the full 413-observation unbalanced panel), reflecting the different sample composition. The decomposition reveals three comparison types with

their respective weights and weighted average estimates: "earlier vs. later treated" (weight = 0.39, estimate = 0.527), "later vs. earlier treated" (weight = 0.40, estimate = 0.398), and "treated vs. untreated" (weight = 0.21, estimate = 0.329). These three components, when weighted by their respective shares, sum to the balanced-panel TWFE of 0.434: $0.39 \times 0.527 + 0.40 \times 0.398 + 0.21 \times 0.329 \approx 0.434$. All three component estimates are of similar magnitude, and the "later vs. earlier" comparisons—which can receive negative weights under treatment effect heterogeneity—have a weighted estimate (0.398) close to the other components, suggesting minimal contamination bias.

C. Robustness Appendix

C.1 Randomization Inference Details

The randomization inference procedure randomly reassigns the vector of treatment-cohort indicators across states 1,000 times, preserving the marginal distribution of adoption timing but breaking the association between treatment timing and state identity. For each permutation, the TWFE specification (Equation (2)) is re-estimated, and the resulting coefficient is stored. The RI p-value is the fraction of permuted coefficients with absolute value at least as large as the observed coefficient.

C.2 Alternative Timing Windows

To assess sensitivity to the precise coding of treatment timing, I shift the assumed law effective date by ±2 years. This tests whether the null result is driven by miscoding the timing of law implementation (e.g., if laws took 2 years to be fully implemented in schools). The results are substantively unchanged under both earlier and later timing assumptions.

D. Severity Gradient

Figure 6 arranges the Sun-Abraham ATT estimates along a severity gradient, from depression (the mildest outcome) through suicide ideation and planning to suicide attempt (the most severe). The SA estimates are: depression 0.074 (SE = 1.520), suicide ideation 0.792 (SE = 1.437), suicide plan 0.873 (SE = 2.176), and suicide attempt 1.170 (SE = 0.574). If anti-cyberbullying laws reduced psychosocial distress, we might expect larger proportional effects on milder outcomes (which have higher prevalence and are more sensitive to marginal improvements in well-being) or on more severe outcomes (if the laws prevented the worst cases). Instead, the point estimates are uniformly positive (in the wrong direction) and none

except the borderline suicide attempt result achieves statistical significance, providing no evidence of beneficial effects along the mental health severity spectrum.

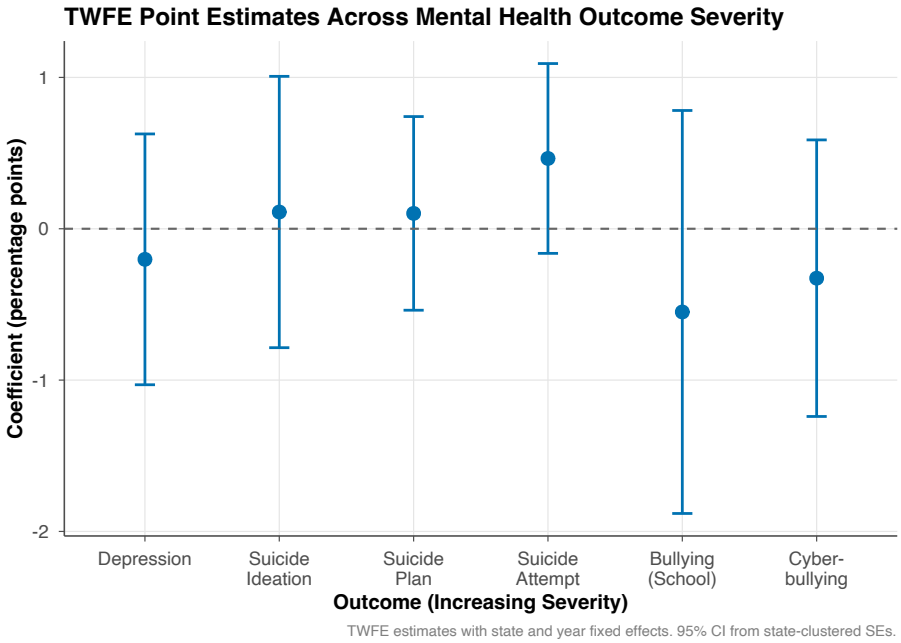

Figure 6: Effect Estimates Across the Mental Health Severity Gradient

E. Heterogeneity Appendix

Additional heterogeneity results are available by grade level (9th–12th) and by race/ethnicity (White, Black, Hispanic). These estimates are omitted from the main text due to small sample sizes within subgroups but are available from the author upon request.