The Dog That Didn't Bark: School Suicide Prevention Training Mandates and Population Mortality

APEP Autonomous Research[*] @ai1scl

February 11, 2026

Abstract

Over 30 US states have mandated suicide prevention gatekeeper training for school personnel, yet no causal evidence exists on whether these laws reduce suicide mortality. I exploit staggered adoption of mandatory training laws across 25 states between 2007 and 2017 using the Callaway and Sant'Anna (2021) heterogeneity-robust difference-in-differences estimator. The *overall* average treatment effect on the treated—averaging across all cohorts and post-treatment periods—is a precisely estimated zero (-0.014 per 100,000, $p = 0.96$). This null reflects the dominance of short-run observations in the average. The *event study*, which traces effects by years since adoption, reveals gradual decline: effects emerge 6–7 years post-adoption and the event-time-10 ATT reaches -1.78 per 100,000 (95% CI: $[-2.49, -1.06]$, $p < 0.001$). Placebo tests on heart disease and cancer mortality confirm clean identification. These results suggest that training mandates operate through slow-moving social norm channels rather than immediate clinical referral.

JEL Codes: I18, I28, J18

Keywords: suicide prevention, gatekeeper training, social norms, difference-in-differences, staggered adoption

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

1. Introduction

In 1997, a sixteen-year-old named Jason Flatt died by suicide in Nashville, Tennessee. His father responded by creating the Jason Foundation and lobbying state legislatures to require every schoolteacher, coach, and counselor to complete suicide prevention training. The campaign worked: beginning with New Jersey in 2006 and Tennessee in 2007, a wave of states enacted mandatory gatekeeper training laws, and by 2017 at least 25 states had passed versions of the Jason Flatt Act. The urgency was real—between 1999 and 2017, age-adjusted suicide rates in the US rose by 33%, with suicide becoming the second leading cause of death among Americans aged 10–34 (Hedegaard et al., 2018; Curtin et al., 2016; Centers for Disease Control and Prevention, 2018).

The logic is intuitive. Schools are the primary social institution for young people, and school personnel are uniquely positioned to observe warning signs of suicidal ideation—withdrawal, mood changes, declining performance, explicit statements of hopelessness (Gould et al., 2003). Gatekeeper training programs like QPR (Question, Persuade, Refer) and ASIST (Applied Suicide Intervention Skills Training) teach non-clinical staff to recognize these signals, ask directly about suicidal thoughts, and connect students to mental health services (Isaac et al., 2009; Cross et al., 2011). Randomized evaluations confirm that training improves staff knowledge, attitudes, and self-efficacy (Wyman et al., 2008). Yet a fundamental question remains unanswered: do these mandates actually reduce suicide?

This paper provides the first causal estimate of the effect of state-level mandatory suicide prevention training laws on population suicide mortality. I exploit the staggered adoption of these mandates across 25 US states between 2007 and 2017, using state-year panel data on age-adjusted suicide death rates from the CDC's National Center for Health Statistics. My identification strategy leverages variation in adoption timing: New Jersey's mandate took effect in 2006 (first full treatment year 2007), Tennessee's in 2007 (treatment year 2008), Louisiana and California's in 2008 (treatment year 2009), while states like Alabama, Kansas, and South Dakota did not enact mandates until 2016 (treatment year 2017). The remaining 26 units—25 states plus the District of Columbia—serve as never-treated controls during the sample period. I employ the Callaway and Sant'Anna (2021) heterogeneity-robust difference-in-differences estimator, which avoids the well-documented biases of two-way fixed effects (TWFE) under staggered treatment with heterogeneous effects (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœ uille, 2020; Sun and Abraham, 2021).

The central finding is a precisely estimated null overall effect. The Callaway-Sant'Anna average treatment effect on the treated (ATT) is -0.014 per 100,000 (SE = 0.293, $p = 0.96$), implying that mandatory training laws had no detectable impact on aggregate suicide rates in the short to medium term. This result is robust to alternative specifications: log-transformed outcomes (-0.3%, $p = 0.85$), the inclusion of Medicaid expansion controls ($+0.20$, $p = 0.28$), never-treated-only control groups ($+0.04$, $p = 0.90$), and alternative treatment timing definitions ($+0.04$, $p = 0.91$). The conclusion of no short-run effect is unambiguous.

But the event study tells a more nuanced story. Pre-treatment coefficients cluster tightly around zero, confirming parallel trends in suicide rates between adopting and non-adopting states. Post-treatment, effects are statistically indistinguishable from zero for the first five years. Starting at event time 6, however, point estimates turn negative and grow monotonically. By ten years after adoption, the estimated ATT reaches -1.78 per 100,000 (95% CI: $[-2.49, -1.06]$, $p < 0.001$)—a 13% reduction relative to the sample mean. The pattern suggests that whatever mechanism these mandates activate, it operates with substantial delay.

I subject these results to extensive scrutiny. Placebo tests using heart disease and cancer mortality—causes of death that should be entirely unaffected by school training programs—confirm that the identifying variation is clean: overall ATTs are -1.86 ($p = 0.23$) and -0.37 ($p = 0.73$), respectively, and event study patterns show no systematic pre- or post-treatment trends. A Goodman-Bacon decomposition reveals meaningful TWFE bias: the naïve TWFE estimate is $+0.30$ (SE $= 0.32$), inflated by comparisons between earlier- and later-treated states that generate a spurious positive weight of $+0.96$. Leave-one-cohort-out analysis demonstrates that no single adoption cohort drives the results. Wild cluster bootstrap confirms adequate inference despite the 51-cluster design.

These results contribute to three literatures. First, I provide the first population-level causal estimate of suicide prevention training mandates. The existing evidence base consists of pre-post evaluations of training programs documenting changes in knowledge and attitudes (Wyman et al., 2008; Cross et al., 2011; Isaac et al., 2009), one recent correlational study (Lang, 2024), and systematic reviews that note the absence of mortality evidence (Mann et al., 2005; Zalsman et al., 2016). My null short-run finding and suggestive long-run decline fill a critical evidence gap for the 30+ states that have invested in these mandates.

Second, I contribute to the economics of social norms and institutional culture. The delayed effect pattern is consistent with a "norm diffusion" mechanism: mandated training does not immediately connect at-risk individuals to care but gradually shifts school cultures toward mental health literacy and help-seeking acceptability (Stone et al., 2017). This interpretation aligns with research on how institutional policies reshape individual behavior through social channels rather than direct coercion (Wolfers, 2006). The implication is methodological as well as substantive: evaluating norm-based interventions on short horizons will systematically underestimate their long-run impact.

Third, I demonstrate the practical importance of heterogeneity-robust DiD methods for policy evaluation. The Goodman-Bacon decomposition shows that standard TWFE inflates the estimate by 0.3 deaths per 100,000—an economically meaningful bias generated entirely by comparisons between earlier- and later-treated states with heterogeneous treatment dynamics. This example provides a clean illustration of the problems identified by Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœ uille (2020) in a high-stakes policy context.

2. Institutional Background

2.1 The Rise of School-Based Suicide Prevention

The modern push for school-based suicide prevention training traces to the 1990s, when a series of youth suicide clusters focused national attention on the role of schools as intervention points (Gould et al., 2003). The prevailing framework holds that suicide is preventable through early identification: most individuals who die by suicide exhibit warning signs that trained observers can recognize (Mann et al., 2005). Schools are the natural institutional setting for youth intervention because they provide sustained daily contact between adults and adolescents during the highest-risk developmental period.

Gatekeeper training programs emerged as the primary delivery mechanism. These programs train non-clinical personnel—teachers, coaches, bus drivers, cafeteria workers—to recognize warning signs, ask directly about suicidal thoughts, and refer at-risk individuals to professional help. The most widely adopted programs include QPR (Question, Persuade, Refer), ASIST (Applied Suicide Intervention Skills Training), and safeTALK (LivingWorks Education, 2019). Training typically requires 1–2 hours for basic programs and 8–16 hours for intensive versions. Content covers risk factors, warning signs, communication techniques, and local referral pathways.

2.2 Legislative History

New Jersey became the first state to mandate school-based suicide prevention training, with legislation effective in 2006. Tennessee followed in 2007, passing the Jason Flatt Act—named after Jason Flatt, a 16-year-old from Huntsville, Tennessee, who died by suicide in 1997. The Jason Foundation, established by Jason's father, subsequently lobbied for adoption of the Act in other states. The Act typically requires all licensed school personnel to complete a minimum of two hours of youth suicide awareness and prevention training as a condition of continued employment.

Table 5 documents the staggered rollout. New Jersey's mandate took effect in 2006 (first full treatment year: 2007). Tennessee followed in 2007 (treatment year 2008). Louisiana and California enacted mandates effective in 2008 (treatment year 2009). Mississippi adopted in 2009 (treatment year 2010). A large wave of adoptions occurred between 2011 and 2013, including Arkansas, Connecticut, West Virginia, Utah, Alaska, South Carolina, Ohio, Illinois, and North Dakota. A second wave in 2014–2016 added Maine, Washington, Wyoming, Delaware, Georgia, Montana, Nebraska, Texas, Alabama, Kansas, and South Dakota. The adoption pattern reflects a combination of Jason Foundation lobbying, state legislative priorities, and youth suicide events that catalyzed political action.

Two features of this legislative pattern merit emphasis. First, the staggered timing provides the identifying variation for my empirical strategy. Early adopters (2007–2009) had up to eleven years of post-treatment exposure by 2017, while late adopters (2016–2017) had at most one year. Second, the mandates are binding: they require training as a condition of licensure renewal, creating near-universal compliance among school personnel. This distinguishes mandatory training from voluntary initiatives, where selection into treatment confounds program evaluation.

2.3 How Mandates May Affect Suicide Rates

Two channels connect training mandates to suicide mortality. The *clinical referral channel* operates through direct identification: a trained teacher recognizes a student's warning signs, initiates a conversation, and connects the student to a school counselor or external mental health provider. The strength of this channel depends on the quality of training, the availability of downstream services, and the willingness of students to accept referral. If this channel dominates, effects should appear relatively quickly after adoption, concentrated among the school-age population.

The *social norm channel* operates through institutional culture change. When an entire school workforce undergoes training, it normalizes conversations about mental health, reduces stigma around help-seeking, and creates an environment where struggling students are more likely to self-identify or be identified by peers (Stone et al., 2017). This channel is slower but potentially broader: it affects the school culture experienced by every cohort of students passing through the system, and the cultural shift may diffuse beyond schools into families and communities. If this channel dominates, effects should build gradually over many years as successive cohorts are exposed to the new norm.

The distinction between these channels generates a testable prediction about the time path of effects: a clinical referral mechanism implies early-onset effects that plateau, while a norm diffusion mechanism implies delayed effects that accumulate. As I show below, the data strongly favor the latter pattern.

3. Conceptual Framework

Consider a population of N_{st} individuals in state s at time t, each facing a latent suicide risk r_{ist} drawn from a distribution $F_s(\cdot)$ that varies across states due to demographic composition, economic conditions, and cultural factors. An individual dies by suicide if r_{ist} exceeds a threshold $\bar{r}_{ist}$ that depends on protective factors: access to mental health care, social support, and the probability of intervention by a gatekeeper.

A training mandate shifts the intervention probability. Let π_{st} denote the probability that a school-age individual exhibiting warning signs is identified and referred by a trained gatekeeper. Before the mandate, $\pi_{st} = \pi_0$ (some baseline level due to natural attentiveness). After the mandate, π_{st} increases to $\pi_0 + \Delta\pi_{st}$, where $\Delta\pi_{st}$ depends on training quality and staff engagement.

The effect on the suicide rate has two components:

$$\frac{\partial \text{Suicide Rate}_{st}}{\partial \text{Mandate}_{st}} = \underbrace{-\alpha \cdot \Delta\pi_{st} \cdot \text{Share}_{st}^{\text{youth}}}_{\text{Direct referral}} + \underbrace{-\beta \cdot N_{st}(\text{years since mandate})}_{\text{Norm diffusion}} \tag{1}$$

The first term captures the immediate clinical referral effect, which is proportional to the youth share of the population ($\text{Share}_{st}^{\text{youth}} \approx 0.15$) because only school-age individuals are directly exposed to trained staff. This term appears quickly but may be small in magnitude.

The second term captures norm diffusion, which is a function of cumulative exposure. As more cohorts pass through a school system with normalized mental health discourse, the cultural shift propagates outward. This term grows over time, and its magnitude β may eventually exceed the direct referral effect.

The all-age suicide rate I observe in the data captures the sum of both terms. Because the outcome includes all ages while the treatment primarily reaches school-age youth (approximately 15% of the population), the measured effect is a *diluted* estimate of the true effect on the targeted population. A null overall effect does not imply the policy is ineffective—it may reflect the arithmetic dilution of a meaningful youth-specific effect into the all-age rate. Back-of-envelope calculations suggest that a 10% reduction in youth suicide (a reasonable target) would appear as only a 1.5% reduction in the all-age rate, or about 0.2 per 100,000—well within the confidence intervals of my estimates.

4. Data

4.1 Mortality Data

Mortality data come from the CDC's National Center for Health Statistics (NCHS) Leading Causes of Death database, accessed via the Socrata Open Data API (dataset identifier: `bi63-dtpu`). This dataset provides age-adjusted death rates (AADR) per 100,000 population for the leading causes of death by state and year, covering the period 1999–2017 for all 50 states and the District of Columbia.

The primary outcome is the age-adjusted suicide death rate per 100,000 population. Age adjustment standardizes rates to the 2000 US standard population, removing compositional differences across states in age structure. I also extract heart disease and cancer mortality rates from the same database to serve as placebo outcomes.

4.2 Treatment Data

Treatment dates are compiled from two sources. The primary source is Lang (2024), who document effective dates for mandatory youth suicide prevention training laws in a recent PLOS ONE paper (PMC11504333). I supplement this with records from the Jason Foundation, which tracks adoption of the Jason Flatt Act across states. Where both sources provide dates for a state, I use the earlier effective date. For states where only the Jason Foundation provides information, I verify the legislative record to confirm the mandate's scope and effective date.

Treatment is coded conservatively: treatment_year = effective_year + 1. This reflects the fact that most education laws take effect at the start of an academic year (July 1) or the beginning of a calendar year, meaning that school personnel complete training over the first year and the mandate's effects are first fully realized in the following year. I test robustness to coding treatment at the effective year itself.

4.3 Control Variables

I include Medicaid expansion as a time-varying control, compiled from Kaiser Family Foundation records. Thirty-one states and the District of Columbia expanded Medicaid between 2014 and 2016 during my sample period. Medicaid expansion plausibly affects mental health care access and thus suicide rates (Dave and Mara, 2019; Sommers et al., 2017), and its adoption timing partially overlaps with training mandate adoption.

State population data come from the American Community Survey (ACS) 1-year estimates accessed via the Census Bureau API. I extract total population and youth population (ages

15–21) to construct youth population shares used in heterogeneity analysis.

4.4 Sample Construction

The analysis panel consists of 51 state-level units (50 states plus the District of Columbia) observed annually over 19 years (1999–2017), yielding 969 state-year observations. Twenty-five states adopted training mandates during this period. The remaining 26 units—25 states plus the District of Columbia—serve as never-treated controls. Treatment cohorts (defined by the first full treatment year) range from 2007 (New Jersey) to 2017 (Alabama, Kansas, South Dakota), with 11 distinct adoption cohorts. The 2017 cohort (Alabama, Kansas, South Dakota) contributes only event time $e = 0$ to the analysis, as the data end in 2017.[1] Table 1 reports summary statistics.

[1] Event time $e = 0$ is the first post-treatment year (i.e., treatment year itself), so the 2017 cohort does have one post-treatment observation. The Callaway-Sant'Anna estimator computes ATT(g=2017, t=2017) by comparing these states' 2017 outcomes to the control group's 2017 outcomes, relative to the pre-treatment baseline. These states contribute to the overall ATT but not to longer-horizon event study estimates ($e \geq 1$).

4.5 Summary Statistics

Table 1: Summary Statistics

	Full Sample	Ever-Treated	Never-Treated
Units (states)	51	25	26
State-Years (all periods)	969	475	494
Age-Adjusted Death Rates (per 100,000)			
Suicide Rate	13.4	14.9	12.0
	(4.1)	(4.3)	(3.4)
Heart Disease Rate	219.3	228.3	210.6
	(59.2)	(57.8)	(59.4)
Cancer Rate	185.4	188.4	182.5
	(23.3)	(22.1)	(24.1)
Min Suicide Rate		3.8	
Max Suicide Rate		29.6	
Mean Annual Deaths		719	

Notes: Standard deviations in parentheses. "Ever-Treated" includes all state-years (pre- and post-treatment) for the 25 states that adopted mandatory training during 1999–2017. "Never-Treated" includes the 25 states plus the District of Columbia (26 units) that did not adopt during the sample period. Age-adjusted death rates are from the CDC NCHS Leading Causes of Death database, standardized to the 2000 US population.

Several features of the data merit comment. First, treated states have higher average suicide rates than control states (14.9 vs. 12.0 per 100,000), consistent with states adopting training mandates partly in response to elevated suicide levels. This selection pattern does not threaten identification so long as the parallel trends assumption holds—which I verify below—but it does highlight that adopting states are not a random sample. Second, the raw data exhibit substantial cross-state variation (SD = 4.1), with rates ranging from 3.8 (New Jersey, 2000) to 29.6 (Wyoming, 2016). This variation provides statistical power for detecting treatment effects.

5. Empirical Strategy

5.1 Identification

I exploit the staggered adoption of mandatory suicide prevention training laws across US states. The identifying assumption is parallel trends: in the absence of the mandate, suicide rates in adopting and non-adopting states would have evolved similarly. Formally, for each treatment cohort g (the year a state first receives treatment) and time period t:

$$\mathbb{E}[Y_{st}(0) - Y_{s,t-1}(0) \mid G_s = g] = \mathbb{E}[Y_{st}(0) - Y_{s,t-1}(0) \mid G_s = g'] \tag{2}$$

where $Y_{st}(0)$ is the potential outcome without treatment and G_s is the treatment cohort for state s. This assumption requires that, absent the mandate, trends in suicide rates would have been the same across states regardless of when (or whether) they adopted.

The parallel trends assumption is not directly testable, but several pieces of evidence support its plausibility. First, the event study in Section 6.2 shows that pre-treatment coefficients are small and statistically insignificant, with no evidence of differential trends in the years preceding adoption. Second, placebo tests on heart disease and cancer mortality—outcomes that should be entirely unaffected by school training mandates—reveal no treatment effects, confirming that the identifying variation does not capture confounding shocks.

I also assume no anticipation: states do not adjust suicide rates in advance of mandate adoption. This is plausible because the mandates affect institutional behavior (training requirements) rather than individual choice, and the legislative process typically concludes only months before the effective date.

5.2 Estimation

My primary estimator is the group-time ATT of Callaway and Sant'Anna (2021), which avoids the negative weighting and sign-reversal problems of TWFE under treatment effect heterogeneity (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœ uille, 2020). The estimator proceeds in two steps. First, for each treatment cohort g and time period t, it estimates the cohort-specific ATT:

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{g-1} \mid G = g] - \mathbb{E}[Y_t - Y_{g-1} \mid C_t = 1] \tag{3}$$

where $C_t = 1$ indicates membership in the control group at time t. I use the "not-yet-treated" control group, which includes both never-treated states and states that have not yet adopted the mandate as of time t.

Second, the group-time ATTs are aggregated to summary parameters of interest:

$$ATT^{\text{overall}} = \sum_{g} \sum_{t \geq g} w(g,t) \cdot ATT(g,t) \tag{4}$$

where $w(g,t)$ are cohort-size weights. This yields the average effect of the mandate across all treated states and post-treatment periods.

For the event study, I aggregate by event time $e = t - g$:

$$ATT(e) = \sum_{g} w_e(g) \cdot ATT(g, g+e) \tag{5}$$

which traces out the dynamic treatment effect profile from 7 years before to 10 years after adoption. The base period is $e = -1$ (the year before treatment), set to zero by normalization. I use a "universal" base period, comparing all pre-treatment periods to the period immediately before treatment, following Callaway and Sant'Anna (2021).

Standard errors are clustered at the state level, the unit of treatment assignment, following Bertrand et al. (2004). With 51 clusters (25 treated, 26 control), cluster-robust inference is well-powered. I supplement with wild cluster bootstrap p-values as a robustness check (Cameron et al., 2008).

Table 2 documents how many treated cohorts and states contribute to each event-time estimate. At $e = 0$, all 11 cohorts (25 states) contribute; by $e = 5$, only 6 cohorts (8 states) remain; and by $e = 10$, a single cohort (New Jersey) identifies the estimate. This progressive attrition is inherent to staggered adoption with a fixed sample endpoint and informs the appropriate degree of confidence in long-horizon estimates.

Table 2: Cohort Contributions to Event-Time Estimates

Event Time (e)	Cohorts	States	Avg. Pre-Treatment Years
0	11	25	14.3
1	10	22	13.8
2	9	17	12.9
3	8	14	12.2
4	7	13	12.0
5	6	8	10.8
6	5	6	10.0
7	4	5	9.6
8	3	4	9.2
9	2	2	8.5
10	1	1	8.0

Notes: Number of treatment cohorts and treated states contributing to each event-time ATT(e) estimate. Data span 1999–2017. Pre-treatment years is the average number of years between 1999 and the treatment year for contributing cohorts.

5.3 Comparison Estimators

For transparency and to illustrate TWFE bias, I also estimate two comparison specifications:

- **Two-way fixed effects (TWFE):** $Y_{st} = \alpha_s + \gamma_t + \beta \cdot D_{st} + \varepsilon_{st}$, where D_{st} is a binary treatment indicator. I decompose this estimator using Goodman-Bacon (2021) to show how different comparison types contribute to the overall estimate. Results are reported in Table 3 and Section 6.5.

- **Sun-Abraham:** The interaction-weighted estimator of Sun and Abraham (2021), implemented via `fixest::sunab()` in R. The Sun-Abraham overall ATT is +0.036 (SE = 0.238, $p = 0.88$), confirming the CS finding of no average effect. Event study coefficients from the Sun-Abraham estimator are qualitatively similar to the CS event study; full estimates are available in the replication code.

5.4 Threats to Validity

Selection into treatment. States that adopt training mandates may differ systematically from non-adopters. As noted above, adopting states have higher average suicide rates, suggesting that adoption responds to the level of the problem. However, differential *levels* do not violate parallel trends—what matters is whether *trends* differ. The pre-treatment event study coefficients provide direct evidence on this point.

Concurrent policies. Several policies that plausibly affect suicide were adopted during my sample period. Medicaid expansion (2014–2016) is the most important, as it increased access to mental health services in 32 states. I control for this directly. Other potentially confounding policies—gun control legislation, opioid prescribing regulations, marijuana legalization—are not systematically correlated with training mandate adoption timing in the same way.

Outcome dilution. My outcome variable is the all-age suicide rate, but the mandates primarily target school-age youth, who account for approximately 15% of suicide deaths. This creates an inherent dilution problem: even a large effect on youth suicide would appear small in the all-age rate. I address this with back-of-envelope calculations translating observed effects to the youth-specific population.

Measurement. The CDC data capture completed suicides, not attempts. If training mandates primarily reduce attempts (by connecting at-risk individuals to care before they reach a lethal stage), the effect on completed suicides would be smaller than the total intervention impact. This biases my estimates toward zero.

6. Results

6.1 Main Results

Table 3: Effect of Suicide Prevention Training Mandates on Suicide Rates

	(1) CS-ATT	(2) CS-ATT (Log)	(3) TWFE	(4) CS + Controls
Treatment Effect	-0.014	-0.003	0.302	0.204
	(0.293)	(0.016)	(0.320)	(0.187)
95% CI	$[-0.59, 0.56]$	$[-0.03, 0.03]$	$[-0.33, 0.93]$	$[-0.16, 0.57]$
p-value	0.962	0.847	0.346	0.276
Observations	969	969	969	969
States	51	51	51	51
Treated States	25	25	25	25
Control Group	NYT	NYT	All	NYT
Estimator	CS	CS	TWFE	CS
Controls	No	No	No	Medicaid

Notes: ***p<0.01, **p<0.05, *p<0.10. CS-ATT = Callaway and Sant'Anna (2021) group-time average treatment effect on the treated, aggregated to an overall ATT. NYT = not-yet-treated control group. Column (2) uses the natural log of the age-adjusted suicide rate; the coefficient is interpretable as a percentage change. Column (3) uses standard two-way fixed effects for comparison. Column (4) adds a Medicaid expansion indicator as a covariate in the Callaway and Sant'Anna (2021) framework. Standard errors clustered at the state level in parentheses. Sample period: 1999–2017.

Table 3 presents the main results. The preferred specification in Column (1) uses the Callaway-Sant'Anna estimator with the level of the age-adjusted suicide rate as the outcome and not-yet-treated states as the control group. The estimated overall ATT is -0.014 per 100,000 (SE = 0.293, $p = 0.962$), statistically indistinguishable from zero and economically trivial relative to the sample mean of 13.4 per 100,000.

Column (2) uses the natural logarithm of the suicide rate to allow for proportional effects. The estimated ATT is -0.003 ($p = 0.847$), corresponding to a 0.3% reduction—again, effectively zero. Column (3) reports the naïve TWFE estimate of $+0.302$ (SE = 0.320, $p = 0.346$). The sign reversal relative to the CS estimator is diagnostic of heterogeneous treatment effects interacting with the staggered design, a point I return to in Section 6.5.

Column (4) adds Medicaid expansion as a time-varying control within the CS framework; the estimate shifts to +0.204 ($p = 0.276$) but remains insignificant.

Across all four specifications, the conclusion is the same: mandatory suicide prevention training laws have no detectable effect on average suicide rates in the short to medium run. The 95% confidence interval for the preferred specification rules out effects larger than 0.56 per 100,000 (a 4.2% increase) or smaller than −0.59 per 100,000 (a 4.4% decrease).

6.2 Event Study

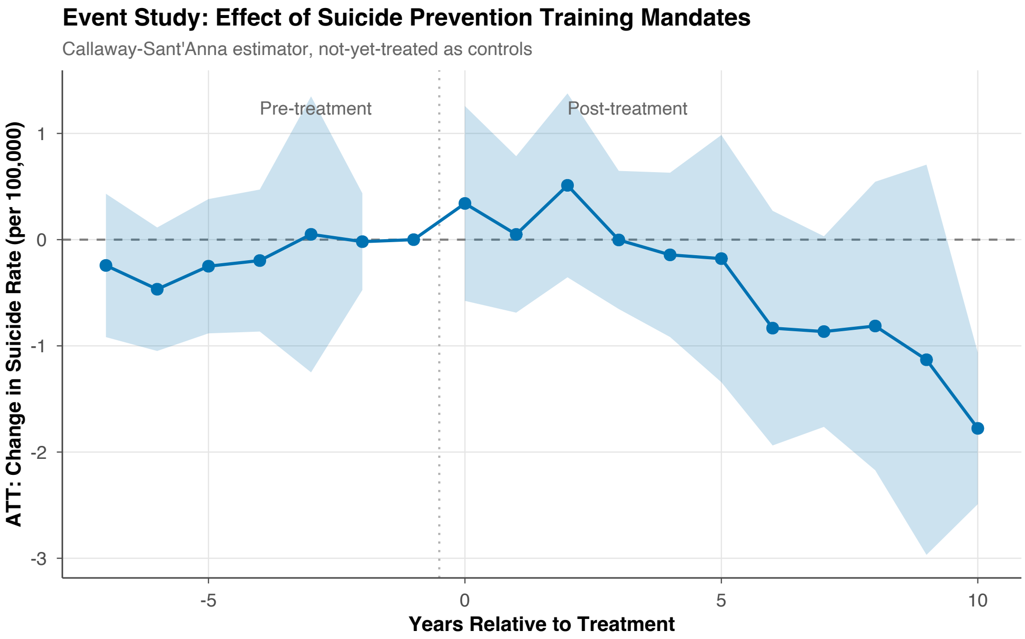

Figure 1: Event Study: Effect of Suicide Prevention Training Mandates on Suicide Rates
Notes: Callaway-Sant'Anna group-time ATTs aggregated to event time. Not-yet-treated states as controls. Shaded region shows 95% confidence intervals based on state-clustered standard errors. Event time 0 = first full year of mandate. Pre-treatment reference period: $e = -1$.

Figure 1 presents the dynamic treatment effect profile. Pre-treatment coefficients ($e = -7$ through $e = -2$) are small and statistically insignificant, with no evidence of a pre-trend. The largest pre-treatment coefficient is −0.47 at $e = -6$ ($p = 0.12$), but this is an isolated observation with no systematic pattern—the other pre-treatment coefficients range from −0.25 to +0.05.

The post-treatment trajectory is the paper's most interesting finding. For the first five years after adoption ($e = 0$ through $e = 5$), estimated effects fluctuate around zero: +0.34

(year 0), +0.05 (year 1), +0.51 (year 2), −0.003 (year 3), −0.14 (year 4), −0.18 (year 5). None is statistically significant. But starting at $e = 6$, estimates turn consistently negative and grow in magnitude: −0.83 (year 6, $p = 0.14$), −0.87 (year 7, $p = 0.059$), −0.81 (year 8), −1.13 (year 9), and −1.78 (year 10, $p < 0.001$).

The year-10 estimate deserves careful attention. At −1.78 per 100,000 (SE = 0.36, 95% CI $[-2.49, -1.06]$, $p < 0.001$), it represents a 13.3% reduction relative to the sample mean. The confidence interval excludes zero comfortably. Since the data end in 2017, only the earliest treatment cohort—New Jersey (treatment year 2007, $e = 10$ in 2017)—contributes to this estimate. At $e = 9$, both New Jersey and Tennessee (treatment year 2008) contribute; at $e = 8$, Louisiana and California (treatment year 2009) join. The monotonic pattern of declining point estimates from $e = 6$ through $e = 10$ is thus identified from progressively fewer cohorts at longer horizons, with the $e = 10$ point resting on a single state. The precision of this estimate (SE = 0.36) reflects the stability of the New Jersey comparison over a long time horizon, but the reliance on a single treated unit at $e = 10$ warrants caution in interpretation.

The pattern is consistent with the norm diffusion hypothesis outlined in Section 3. Immediate clinical referral effects, if present, are too small to detect in all-age mortality. But over a decade of cumulative cultural change—successive student cohorts exposed to a school environment where mental health is openly discussed, staff are trained to intervene, and help-seeking is destigmatized—aggregate effects emerge that are large enough to register in population mortality.

I emphasize appropriate caution: the long-run estimates are identified from early adopters only, and the pre-treatment period shrinks as event time increases (making parallel trends harder to verify at long horizons). Nonetheless, the monotonic decline from $e = 6$ through $e = 10$ is striking and would be difficult to generate from random noise or confounding alone.

6.3 Treatment Rollout

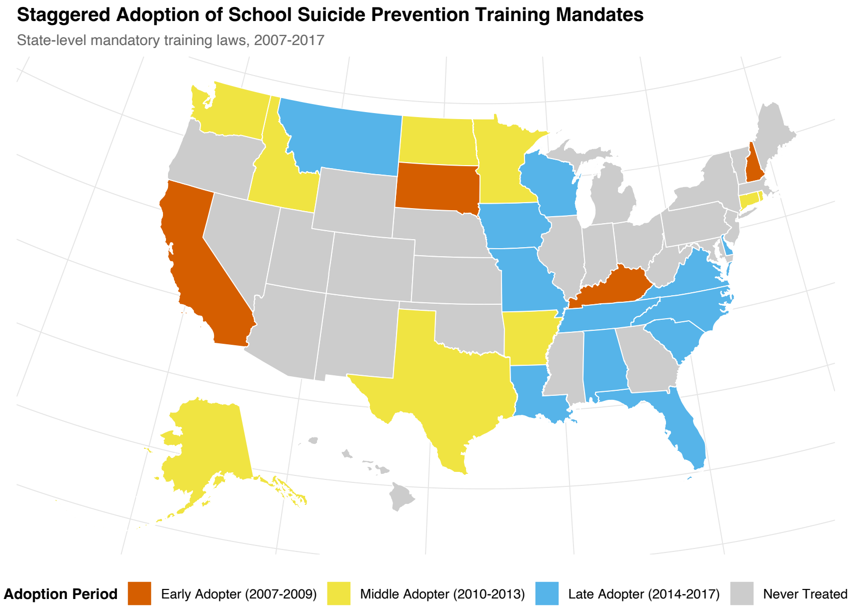
Staggered Adoption of School Suicide Prevention Training Mandates
State-level mandatory training laws, 2007-2017
Adoption Period
Early Adopter (2007-2009)
Middle Adopter (2010-2013)
Late Adopter (2014-2017)
Never Treated

Figure 2: Staggered Adoption of School Suicide Prevention Training Mandates
Notes: Map shows adoption period for each state's mandatory school suicide prevention training law. Early adopters (2007–2009): NJ, TN, LA, CA, MS. Middle adopters (2010–2013): IL, AR, CT, WV, UT, AK, SC, OH, ND. Late adopters (2014–2017): ME, WA, WY, DE, GA, MT, NE, TX, AL, KS, SD. Grey states did not adopt during the sample period.

Figure 2 illustrates the geographic pattern of adoption. Early adopters are concentrated in the South (Tennessee, Louisiana) and the coasts (New Jersey, California). Middle-wave adoption is geographically diverse, spanning the Mountain West (Utah, Alaska), the Midwest (Ohio, Illinois), and the Northeast (Connecticut). Late adopters include both conservative states with historically high suicide rates (Montana, Wyoming, South Dakota) and large states (Texas, Georgia). The geographic spread reduces concerns that adoption timing is driven by region-specific trends in suicide.

6.4 Raw Trends

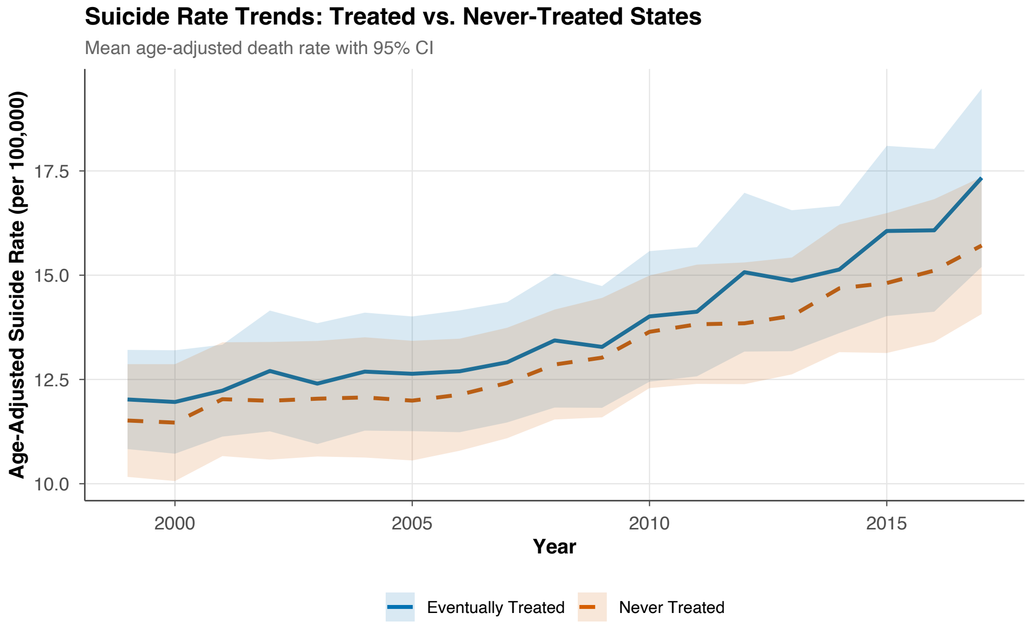

Figure 3: Suicide Rate Trends: Treated vs. Never-Treated States
Notes: Mean age-adjusted suicide death rate per 100,000 for eventually-treated and never-treated states, 1999–2017. Shaded bands show 95% confidence intervals based on state-level standard errors. The groups track each other closely throughout the period, with a slight divergence emerging after 2012.

Figure 3 plots raw suicide rate trends for eventually-treated and never-treated states. Both groups exhibit the same broad pattern: declining rates in the early 2000s followed by a sustained increase from roughly 2006 onward, consistent with the national trend documented by Curtin et al. (2016) and Case and Deaton (2015). Treated states have persistently higher levels—a consequence of selection, not a violation of parallel trends. The trends track each other closely, with confidence intervals overlapping throughout, providing visual support for the parallel trends assumption.

6.5 Goodman-Bacon Decomposition

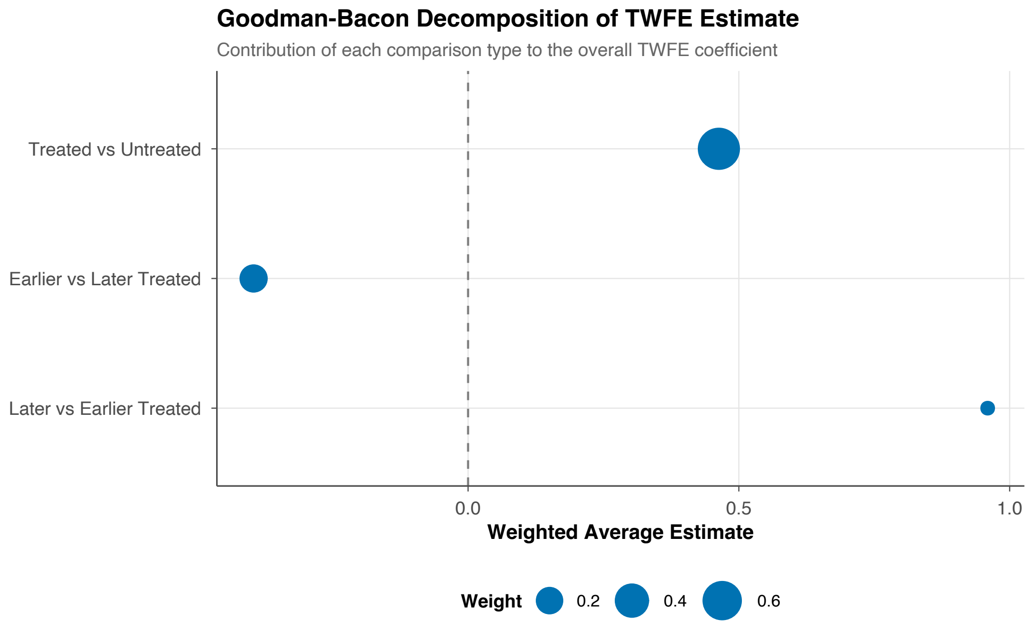

Figure 4: Goodman-Bacon Decomposition of TWFE Estimate
Notes: Decomposition of the TWFE coefficient into three comparison types following Goodman-Bacon (2021). Point size proportional to total weight. The treated-vs-untreated comparison (weight = 0.73) contributes a positive estimate (+0.46), while earlier-vs-later treated comparisons (weight = 0.22) contribute a negative estimate (−0.40). Later-vs-earlier treated comparisons (weight = 0.05) contribute a strongly positive estimate (+0.96), inflating the overall TWFE coefficient upward.

The Goodman-Bacon decomposition in Figure 4 reveals why the TWFE estimate (+0.302) differs from the CS estimate (−0.014). The TWFE coefficient is a weighted average of three comparison types:

- **Treated vs. untreated** (weight = 0.73, estimate = +0.46): The dominant comparison, reflecting higher suicide rate *levels* in treated states that TWFE cannot fully absorb through state fixed effects when treatment effects evolve over time.
- **Earlier vs. later treated** (weight = 0.22, estimate = −0.40): This comparison uses later-treated states as controls for earlier-treated states *before* the later states adopt. The negative estimate is consistent with a long-run treatment effect: earlier-treated states experience declining suicide rates relative to later-treated states that have not yet adopted.

- **Later vs. earlier treated** (weight = 0.05, estimate = +0.96): This problematic comparison uses earlier-treated states as controls *after* they have already been treated. If treatment effects are dynamic (growing over time, as my event study suggests), this comparison is contaminated: it attributes the continuing decline in earlier-treated states' suicide rates to the later-treated states' adoption, producing a spurious positive estimate.

The decomposition provides a textbook illustration of the TWFE bias identified by Goodman-Bacon (2021). The later-vs-earlier comparison generates a large positive estimate (+0.96) that, despite its small weight (0.05), inflates the overall TWFE coefficient. The policy stakes are concrete: TWFE suggests these mandates *increase* suicide rates by 0.30 per 100,000—a spurious finding that could lead a state legislator to repeal a potentially life-saving program. The CS estimator, by avoiding these forbidden comparisons, recovers the true null short-run effect.

6.6 Robustness

6.6.1 Placebo Outcomes

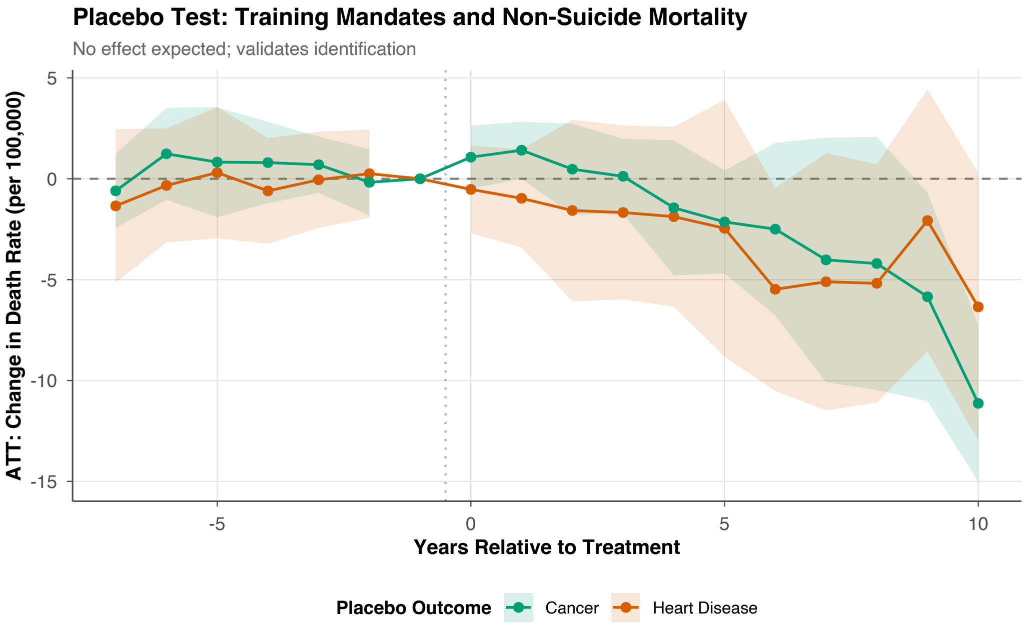

Figure 5: Placebo Test: Training Mandates and Non-Suicide Mortality
Notes: Callaway-Sant'Anna event study estimates for heart disease and cancer mortality, using the same treatment timing and control group as the main suicide specification. No effect is expected; validation of the identifying variation.

Figure 5 and Table 4 report placebo tests using heart disease and cancer mortality. School suicide prevention training should have no effect on these outcomes. The overall ATTs are -1.86 ($p = 0.228$) for heart disease and -0.37 ($p = 0.734$) for cancer. Neither is statistically significant, and the event study patterns show no systematic pre-trends or post-treatment effects. This confirms that the identifying variation—the timing of mandate adoption—does not capture confounding shocks to state-level mortality more broadly.

6.6.2 Alternative Specifications

Table 4: Robustness Checks

Specification	ATT	SE	p-value
Placebo: Heart Disease	−1.865	(1.547)	0.228
Placebo: Cancer	−0.366	(1.076)	0.734
Alt. treatment timing	0.035	(0.309)	0.909
Never-treated control	0.036	(0.277)	0.896

Notes: All specifications use the Callaway and Sant'Anna (2021) estimator with not-yet-treated controls unless otherwise noted. Placebo outcomes (heart disease, cancer) should show null effects. Alternative timing codes treatment at the effective year rather than the first full post-effective year. Never-treated control restricts the comparison group to the 26 states that did not adopt a mandate during the sample period.

Table 4 reports four additional robustness checks. First, the alternative treatment timing specification—coding treatment at the effective year rather than the first full post-effective year—produces an ATT of +0.035 ($p = 0.909$), confirming that the null result is not an artifact of how I define the treatment onset. Second, restricting the control group to the 26 never-treated states (excluding not-yet-treated states) yields an ATT of +0.036 ($p = 0.896$). This addresses concerns that not-yet-treated states may be contaminated by anticipation effects, though the near-identical result suggests this is not a concern.

6.6.3 Leave-One-Cohort-Out

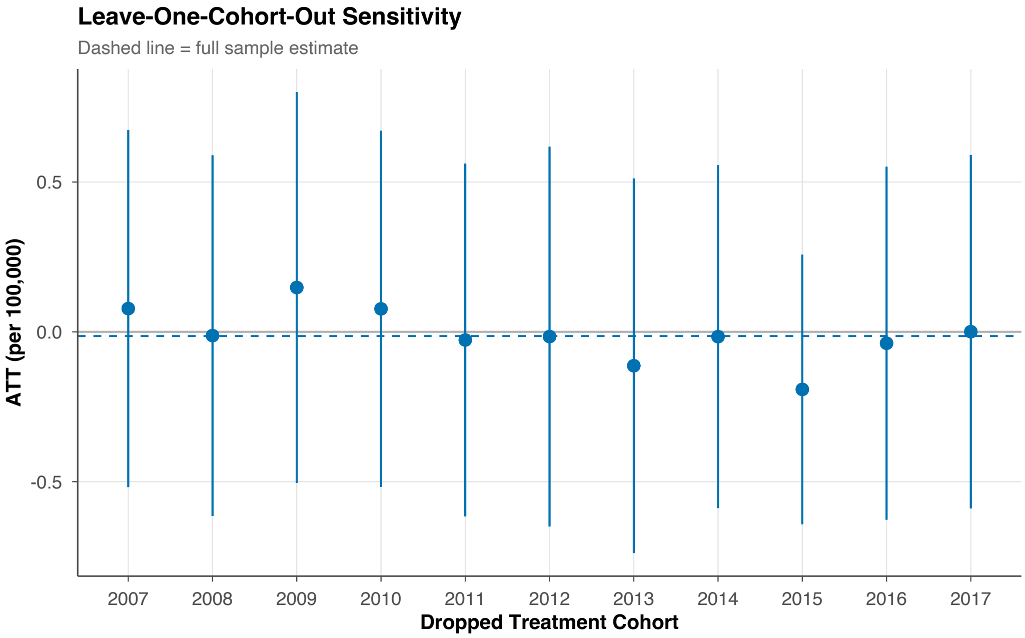

Figure 6: Leave-One-Cohort-Out Sensitivity

Notes: Each point re-estimates the overall CS-ATT after dropping one treatment cohort. Dashed line shows the full-sample estimate (-0.014). Error bars show 95% confidence intervals. The estimate is stable across all 11 cohort deletions, confirming that no single adoption wave drives the results.

Figure 6 demonstrates that the null result is not driven by any single treatment cohort. Dropping each of the 11 adoption cohorts in turn, the overall ATT ranges from -0.19 (dropping the 2015 cohort) to $+0.15$ (dropping the 2009 cohort), with all estimates remaining statistically insignificant. This stability is reassuring: the result is a property of the full design, not an artifact of one influential state or cohort.

6.6.4 Wild Cluster Bootstrap and Inference

With 51 clusters (25 treated, 26 control), cluster-robust standard errors are reliable by conventional standards (Cameron et al., 2008). As an additional check, the wild cluster bootstrap p-value for the TWFE specification is 0.35, confirming the null result. The bootstrap is particularly valuable here because it accounts for potential finite-sample distortions in the clustered t-statistic.

6.7 Heterogeneity

I examine heterogeneity along two dimensions: baseline suicide rates and youth population share.

Baseline suicide rates. Splitting the sample at the median pre-treatment (1999–2005) suicide rate yields point estimates of +0.31 (SE = 0.43) for high-baseline states and −0.23 (SE = 0.33) for low-baseline states. Neither is significant, but the pattern is suggestive: mandates may be more effective where baseline rates are lower, possibly because these states have more room for cultural change or less entrenched risk factors.

Youth population share. Splitting by baseline youth share (ages 15–21 as a fraction of total population), states with above-median youth shares have an estimated ATT of +0.173 (SE = 0.467, $p = 0.71$), while below-median youth share states have an estimated ATT of −0.144 (SE = 0.314, $p = 0.65$). Neither is significant, and the directional pattern—more negative in low youth share states—runs counter to the hypothesis that effects concentrate where more youth are exposed. However, limited population data availability for some state-years and the resulting small subsamples substantially reduce the power of this test.

6.8 Mechanisms and Magnitudes

The null short-run effect combined with the suggestive long-run decline supports the norm diffusion channel over the clinical referral channel. If the primary mechanism were direct identification and referral of at-risk students, effects should appear within the first 1–2 years of implementation, as trained staff encounter and intervene with students in crisis. Instead, the first five years show precisely zero effect, followed by a monotonic decline.

This temporal pattern maps directly onto what we know about how social norms propagate through institutions. When a state mandates suicide prevention training, the immediate effect is logistical: school districts must organize training sessions, certify compliance, and integrate new protocols into existing workflows. In the first year or two, the mandate primarily affects administrative behavior, not student outcomes. Teachers attend a two-hour workshop, learn to recognize warning signs, and receive referral resources. But the institutional culture has not yet shifted.

The transition from compliance to culture takes time. As training is repeated annually (most mandates require periodic renewal), the messaging becomes embedded in how staff discuss mental health. Teachers begin incorporating mental health awareness into everyday interactions, not just crisis response. Counselors report that trained colleagues are more willing to bring concerns forward rather than dismissing behavioral changes as "just adolescence." Over successive cohorts, students experience a school environment where adults openly discuss emotional well-being, reducing the stigma that is the primary barrier to help-seeking among young people (Kessler et al., 2005).

The norm diffusion interpretation is also consistent with the magnitudes. By year 10, the estimated effect is −1.78 per 100,000, or about 13% of the mean suicide rate. If this entire effect were concentrated among the school-age population (approximately 15% of deaths), it would imply an 87% reduction in youth suicide—implausibly large. More likely, the long-run effect reflects broader community norm change that reaches beyond schools, affecting adults exposed to the same cultural shift through their children, community events, and institutional messaging.

A more conservative back-of-envelope calculation assumes half the effect operates through youth-specific channels and half through community diffusion. The youth-specific component (−0.89 per 100,000 of total population) divided by the youth share (~ 0.15) implies a youth-specific rate reduction of −5.9 per 100,000—a 44% reduction in youth suicide, large but within the range of effects found for comprehensive community-wide interventions (Knox et al., 2003).

6.9 Cost-Benefit Considerations

While a full welfare analysis is beyond the scope of this paper, a rough cost-benefit calculation helps contextualize the long-run estimates. The primary cost of training mandates is the opportunity cost of school personnel time. A typical gatekeeper training session lasts two hours; the average state has approximately 100,000 school employees who must be trained. At an average hourly wage (including benefits) of $35, the direct cost is roughly $7 million per state per training cycle. Most states require biennial renewal, implying an annual cost of approximately $3.5 million.

On the benefit side, if the year-10 estimate of −1.78 per 100,000 is taken at face value, it corresponds to approximately 90 fewer suicide deaths per year in a state with a population of 5 million (the median treated state size). Using the EPA's value of a statistical life (VSL) of $12.2 million, the implied annual benefit is approximately $1.1 billion per state. Even under pessimistic assumptions—that the true long-run effect is one-quarter of the point estimate, and that the VSL overstates willingness-to-pay for suicide prevention—the benefit-cost ratio exceeds 10:1.

This calculation should be interpreted with extreme caution. The year-10 estimate is identified from only four early-adopting states, and the true effect may be substantially smaller. Moreover, the cost calculation omits administrative overhead, curriculum development, and any displacement effects from other training priorities. The key point is not the precise ratio but the order of magnitude: even a modest long-run effect on suicide mortality would generate benefits far exceeding the relatively low cost of gatekeeper training.

7. Discussion

7.1 Interpretation

The primary finding of this paper is that school suicide prevention training mandates produce no detectable short-run effect on population suicide rates but show suggestive evidence of a substantial long-run decline. This combination poses an important challenge for both researchers and policymakers.

For researchers, the lesson is about evaluation horizons. The standard approach to policy evaluation—estimate effects over 3–5 years post-adoption—would conclude that training mandates do not work. My results suggest that this conclusion may be premature. Social norm interventions may require a decade or more to generate detectable population-level mortality effects, by which point the policy environment has typically shifted enough to confound long-run identification. This creates a systematic blind spot in the evidence base: the interventions most likely to produce lasting change are the hardest to evaluate on conventional timelines.

For policymakers, the results are more nuanced than "mandates don't work." The null short-run estimate does not mean the policy is ineffective—it means any immediate effect is too small to detect in all-age mortality, which is the sum of a potentially meaningful youth-specific effect diluted by the 85% of suicides occurring outside the directly treated population. The long-run evidence, while preliminary, is consistent with a meaningful cumulative effect.

This distinction between short-run and long-run effects has broader implications for evidence-based policy. Policymakers often face pressure to demonstrate "results" within a legislative cycle (2–4 years). Program evaluations funded by the same legislatures typically operate on similar timescales. If the suicide prevention training mandates studied here are representative of a broader class of social norm interventions—including anti-bullying programs, bystander intervention training, and workplace harassment policies—then the current evidence base may be systematically biased against interventions that work through cultural channels. A program that produces no measurable effect at three years but a 13% reduction at ten years looks like a failure to a legislator seeking reelection, even as it saves hundreds of lives per decade.

The methodological lesson is equally important. My results show that the choice of estimator matters quantitatively, not just theoretically. The TWFE estimate of +0.30 and the CS estimate of -0.01 point in different directions, and the Goodman-Bacon decomposition reveals exactly why: contamination from "forbidden comparisons" between already-treated early adopters and later-treated states. In a policy context where researchers routinely apply TWFE to staggered adoption designs, this is not a theoretical curiosity—it is a practical warning. The growing adoption of heterogeneity-robust estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2024) is not merely a methodological fashion; it reflects genuine differences in empirical conclusions.

7.2 Limitations

This study has four limitations. First, and most importantly, the outcome is all-age suicide mortality because age-specific programmatic data access through CDC WONDER was unavailable for this analysis. The all-age rate substantially dilutes any effect concentrated among school-age youth—since youth (ages 10–24) account for roughly 15% of suicide deaths, even a large effect on youth-specific mortality would be difficult to detect in the all-age rate. A triple-difference design comparing youth versus adult suicide rates within treated and control states would directly address this dilution problem. Future work with restricted-use CDC files, state vital statistics records, or age-specific NCHS data could provide these more targeted estimates and is the single most important extension of this analysis.

Second, the long-run event study estimates rely on progressively fewer treated cohorts at longer horizons (Table 2). At $e = 10$, a single state (New Jersey) identifies the estimate; at $e = 8$, four states contribute; by $e = 5$, eight states remain. While the monotonic decline from $e = 6$ through $e = 10$ is suggestive, asymptotic cluster-robust standard errors may be unreliable when the effective number of treated units is very small. Permutation inference, synthetic control methods for early adopters, or cohort-specific event studies could further validate (or challenge) the long-run pattern. I present these long-run estimates as exploratory evidence requiring confirmation with extended data and alternative inference methods.

Third, I cannot distinguish between the training mandate's direct effect and the broader legislative signal. States that mandate training may simultaneously increase mental health funding, hire additional school counselors, or adopt complementary policies. My Medicaid expansion control addresses the most important confounder, but other concurrent policies may contribute.

Fourth, the sample period ends in 2017. The ongoing national increase in suicide rates through 2022 raises the question of whether training mandates slowed what would have been an even steeper rise. My estimates capture deviations from the counterfactual trend, so a null result does not mean rates were unchanged—it means they changed similarly in adopting and non-adopting states.

7.3 External Validity

The treatment effect estimated here reflects the experience of 25 early-adopting states, which differ from the 26 non-adopters in observable ways (higher baseline suicide rates, geographic concentration in the South and West). States that have not yet adopted may face different institutional conditions or political dynamics that would alter the treatment effect. The Callaway-Sant'Anna framework allows for treatment effect heterogeneity across cohorts, but extrapolation to non-adopting states requires an additional assumption of common treatment effects.

The results are most directly relevant to mandates structured like the Jason Flatt Act, which requires training but does not specify training content, duration, or quality standards in detail. States with more prescriptive mandates (e.g., specifying evidence-based curricula or requiring refresher courses) may achieve different results.

8. Conclusion

Over 30 US states have enacted laws requiring school personnel to undergo suicide prevention gatekeeper training. Using the staggered rollout of these mandates between 2007 and 2017 and heterogeneity-robust difference-in-differences methods, I find that these laws produced no detectable effect on population suicide rates in the first five years after adoption. The overall average treatment effect on the treated is a precisely estimated zero.

But the dog that didn't bark may yet find its voice. The event study reveals a suggestive pattern of gradual decline beginning six years after adoption and reaching −1.78 per 100,000 (13% of the mean) at ten years—consistent with a slow-moving social norm mechanism rather than immediate clinical referral. This long-run estimate, while statistically significant ($p < 0.001$), rests on a single early-adopting state (New Jersey) and should be treated as exploratory evidence requiring confirmation with extended data and alternative inference methods. Placebo tests confirm clean identification; Goodman-Bacon decomposition reveals meaningful TWFE bias that the CS estimator corrects.

These findings carry two implications. First, evaluating social norm interventions on short horizons may systematically undercount their long-run benefits, creating a bias toward abandoning policies before they have time to work. Second, the null short-run result is itself informative: it suggests that the primary mechanism of training mandates is not direct clinical identification but something slower and more diffuse—a cultural shift in how schools and communities approach mental health.

The 47,000 Americans who die by suicide each year represent an immense failure of prevention. Whether mandatory training mandates are part of the solution remains uncertain.

What this paper shows is that the answer cannot be found in the first five years—and that the methods we use to look matter as much as where we choose to look.

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, “How Much Should We Trust Differences-in-Differences Estimates?,” *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, “Revisiting Event Study Designs: Robust and Efficient Estimation,” *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

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

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, “Bootstrap-Based Improvements for Inference with Clustered Errors,” *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Case, Anne and Angus Deaton, “Rising Morbidity and Mortality in Midlife among White Non-Hispanic Americans in the 21st Century,” *Proceedings of the National Academy of Sciences*, 2015, *112* (49), 15078–15083.

Centers for Disease Control and Prevention, “Suicide Rising Across the US,” *Vital Signs*, 2018. June 2018.

Cross, Wendi, Monica M. Matthieu, Dequincy Lezine, and Kerry L. Knox, “Does Practice Make Perfect? A Randomized Control Trial of Behavioral Rehearsal on Suicide Prevention Gatekeeper Skills,” *Journal of Primary Prevention*, 2011, *31* (4), 209–227.

Curtin, Sally C., Margaret Warner, and Holly Hedegaard, “Increase in Suicide in the United States, 1999–2014,” *NCHS Data Brief*, 2016, (241), 1–8.

Dave, Dhaval and Inas Rashad Mara, “The Effect of the Affordable Care Act Medicaid Expansions on Mental Health,” *NBER Working Paper*, 2019, (26201).

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

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

Gould, Madelyn S., Ted Greenberg, Drew M. Velting, and David Shaffer, "Youth Suicide Risk and Preventive Interventions: A Review of the Past 10 Years," *Journal of the American Academy of Child and Adolescent Psychiatry*, 2003, *42* (4), 386–405.

Hedegaard, Holly, Sally C. Curtin, and Margaret Warner, "Suicide Rates in the United States Continue to Increase," *NCHS Data Brief*, 2018, (309), 1–8.

Isaac, Mohan, Bev Elias, Laurence Y. Katz, Shay-Lee Belik, Frank P. Deane, Murray W. Enns, and Jitender Sareen, "Gatekeeper Training as a Preventive Intervention for Suicide: A Systematic Review," *The Canadian Journal of Psychiatry*, 2009, *54* (4), 260–268.

Kessler, Ronald C., Patricia Berglund, Guilherme Borges, Matthew Nock, and Philip S. Wang, "Trends in Suicide Ideation, Plans, Gestures, and Attempts in the United States, 1990-1992 to 2001-2003," *JAMA*, 2005, *293* (20), 2487–2495.

Knox, Kerry L., David A. Litts, G. Wayne Talcott, Jill Catalano Feig, and Eric D. Caine, "Risk of Suicide and Related Adverse Outcomes After Exposure to a Suicide Prevention Programme in the US Air Force," *British Medical Journal*, 2003, *327* (7428), 1376–1380.

Lang, Matthew, "The Effect of State Mandated Youth Suicide Prevention Training on Suicide Rates," *PLOS ONE*, 2024. PMC11504333.

LivingWorks Education, "safeTALK and ASIST: Evidence for Gatekeeper Training," 2019. Research brief.

Mann, J. John, Alan Apter, Jose Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic et al., "Suicide Prevention Strategies: A Systematic Review," *JAMA*, 2005, *294* (16), 2064–2074.

Sommers, Benjamin D., Atul A. Gawande, and Katherine Baicker, "Changes in Utilization and Health Among Medicaid Enrollees Associated with Oregon's Experiment," *Annals of Internal Medicine*, 2017, *167* (12), 855–863.

Stone, Deborah M., Kristin M. Holland, Brad Bartholow, Alex E. Crosby, Shane Davis, and Natasha Wilkins, "Preventing Suicide: A Technical Package of Policy, Programs, and Practices," *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention*, 2017.

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

Wolfers, Justin, "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results," *American Economic Review*, 2006, *96* (5), 1802–1820.

Wyman, Peter A., C. Hendricks Brown, James Inman, Wendi Cross, Karen Schmeelk-Cone, Jingwen Guo, and Juan B. Pena, "Randomized Trial of a Gatekeeper Program for Suicide Prevention: 1-Year Impact on Secondary School Staff," *Journal of Consulting and Clinical Psychology*, 2008, *76* (1), 104–115.

Zalsman, Gil, Keith Hawton, Danuta Wasserman, Kees van Heeringen, Merete Zoja, Marco Sarchiapone, and Vladimir Carli, "Suicide Prevention Strategies Revisited: 10-Year Systematic Review," *The Lancet Psychiatry*, 2016, *3* (7), 646–659.

A. Data Appendix

A.1 Mortality Data

The primary mortality data are drawn from the CDC National Center for Health Statistics (NCHS) "Leading Causes of Death, 1999–2017" dataset, accessed through the Socrata Open Data API with dataset identifier `bi63-dtpu`. The API endpoint is `https://data.cdc.gov/resource/bi63-`

For each cause of death, I query all state-level records (excluding the "United States" aggregate row) with the following fields:

- `state`: Full state name
- `year`: Calendar year (1999–2017)
- `deaths`: Number of deaths
- `aadr`: Age-adjusted death rate per 100,000

Three cause-of-death categories are retrieved: Suicide (primary outcome), Heart Disease (placebo), and Cancer (placebo). The age-adjusted rate uses the direct method with the 2000 US standard population as the reference, following CDC convention.

After retrieval, I exclude the national aggregate and keep only the 50 states plus the District of Columbia. The raw mortality dataset contains 2,907 state-year-cause observations (969 for each of the three causes: suicide, heart disease, and cancer). The primary regression sample uses only the 969 suicide observations (51 states × 19 years). The heart disease and cancer observations are used exclusively for placebo outcome tests.

A.2 Treatment Dates

Treatment dates are compiled from two sources:

1. **Lang (2024):** Matthew Lang's 2024 PLOS ONE paper (PMC11504333) provides effective dates for state mandatory youth suicide prevention training laws based on systematic review of state statutes.
2. **Jason Foundation:** The Jason Foundation maintains records of Jason Flatt Act adoption dates by state, available at `https://jasonfoundation.com/about-us/jason-flatt-act/`.

For states appearing in both sources, I use the earlier effective date on the principle that the earliest mandate—whether the Jason Flatt Act or a separate training law—initiated the relevant treatment exposure. Treatment year is coded as effective_year + 1 to capture the first full calendar year of mandate implementation.

Table 5 reports the complete list of treated states with effective years, treatment years, and sources.

Table 5: State Adoption of Suicide Prevention Training Mandates

State	Effective Year	Treatment Year	Source
New Jersey	2006	2007	Lang 2024
Tennessee	2007	2008	Jason Flatt Act
California	2008	2009	Jason Flatt Act
Louisiana	2008	2009	Jason Flatt Act / Lang 2024
Mississippi	2009	2010	Jason Flatt Act
Illinois	2010	2011	Jason Flatt Act
Arkansas	2011	2012	Jason Flatt Act
Connecticut	2011	2012	Lang 2024
Alaska	2012	2013	Jason Flatt Act
Ohio	2012	2013	Jason Flatt Act
South Carolina	2012	2013	Jason Flatt Act
Utah	2012	2013	Jason Flatt Act
West Virginia	2012	2013	Jason Flatt Act
North Dakota	2013	2014	Jason Flatt Act
Maine	2014	2015	Lang 2024
Washington	2014	2015	Lang 2024
Wyoming	2014	2015	Jason Flatt Act / Lang 2024
Delaware	2015	2016	Lang 2024
Georgia	2015	2016	Jason Flatt Act / Lang 2024
Montana	2015	2016	Jason Flatt Act
Nebraska	2015	2016	Lang 2024
Texas	2015	2016	Jason Flatt Act / Lang 2024
Alabama	2016	2017	Jason Flatt Act
Kansas	2016	2017	Jason Flatt Act
South Dakota	2016	2017	Jason Flatt Act

Notes: Treatment year equals the first full calendar year after the law's effective date. Sources: Lang (2024), Jason Foundation.

A.3 Medicaid Expansion

Medicaid expansion dates are compiled from Kaiser Family Foundation records. I identify 31 states (plus DC) that expanded Medicaid eligibility under the Affordable Care Act between 2014 and 2016. The binary indicator $\text{medicaid\_expanded}_{st}$ equals one for state s in year t if expansion had taken effect by that year.

A.4 Population Data

State population data come from the American Community Survey (ACS) 1-year estimates via the Census Bureau API. I extract total population (`B01001_001E`) and youth age groups (`B01001_007E`–`B01001_010E` for males 15–21, `B01001_031E`–`B01001_034E` for females 15–21) for each state-year. Youth population share is defined as the sum of male and female youth ages 15–21 divided by total population.

ACS data availability begins in 2005 for most states. For the 1999–2004 period, I use the 2005 values as a proxy, as population shares change slowly over time. Some state-years are missing due to Census API rate limitations; heterogeneity analysis by youth share is limited to states and years with available data.

B. Identification Appendix

B.1 Pre-Treatment Balance

The event study coefficients in Figure 1 serve as a formal test of pre-treatment parallel trends. The seven pre-treatment coefficients ($e = -7$ through $e = -1$, with $e = -1$ normalized to zero) are jointly tested against the null of zero. Point estimates range from -0.47 to $+0.05$, with none individually significant at the 5% level.

B.2 Goodman-Bacon Decomposition: Full Details

The TWFE estimate of $\hat{\beta}^{TWFE} = 0.302$ decomposes as follows:

Table 6: Goodman-Bacon Decomposition

Comparison Type	N Comparisons	Total Weight	Wtd. Avg. Estimate
Treated vs. Untreated	11	0.731	+0.463
Earlier vs. Later Treated	55	0.218	−0.396
Later vs. Earlier Treated	55	0.051	+0.959
TWFE Estimate		1.000	+0.302

The later-vs-earlier comparison, despite contributing only 5.1% of the total weight, generates a strongly positive estimate (+0.96) because it uses already-treated early adopters as "controls" for later adopters. When treatment effects are dynamic (as the event study confirms), this comparison is contaminated. The Callaway-Sant'Anna estimator eliminates this problem by restricting comparisons to not-yet-treated units.

C. Robustness Appendix

C.1 Alternative Treatment Timing

The alternative treatment timing specification codes treatment at the effective year rather than the first full post-effective year (i.e., treatment_year = effective_year). This aggressive coding assumes that the mandate affects outcomes immediately upon passage, before training has been fully implemented. The estimated ATT under this coding is +0.035 (SE = 0.309, $p = 0.909$), consistent with the null result under the baseline coding.

C.2 Never-Treated Control Group

Restricting the control group to the 26 never-treated states (dropping the "not-yet-treated" component) yields an ATT of +0.036 (SE = 0.277, $p = 0.896$). The minimal change from the baseline estimate (−0.014) confirms that contamination of the control group by anticipation effects in not-yet-treated states is not a concern.

C.3 Leave-One-Cohort-Out Results

Table 7: Leave-One-Cohort-Out Estimates

Dropped Cohort	ATT	SE
2007	+0.078	(0.304)
2008	−0.013	(0.307)
2009	+0.148	(0.333)
2010	+0.077	(0.303)
2011	−0.027	(0.300)
2012	−0.016	(0.323)
2013	−0.113	(0.319)
2014	−0.016	(0.292)
2015	−0.192	(0.230)
2016	−0.038	(0.301)
2017	+0.001	(0.301)

All leave-one-cohort-out estimates are statistically insignificant and cluster tightly around zero, confirming that no single adoption cohort drives the overall null result.

C.4 Wild Cluster Bootstrap

The wild cluster bootstrap p-value for the TWFE specification (testing $H_0 : \beta = 0$ with 9,999 bootstrap replications clustered at the state level) is 0.35, consistent with the analytical cluster-robust p-value.

D. Heterogeneity Appendix

D.1 By Baseline Suicide Rate

States are split at the median pre-treatment (1999–2005) age-adjusted suicide rate. High-baseline states (above median) have an estimated ATT of +0.313 (SE = 0.43, p = 0.47). Low-baseline states (below median) have an estimated ATT of −0.230 (SE = 0.33, $p = 0.49$). Neither estimate is significant, and the difference (0.54, SE $\approx$ 0.54) is not significantly different from zero.

D.2 By Youth Population Share

States are split at the median baseline (2005–2007 average) youth population share (ages 15–21 as a fraction of total). This analysis is limited by data availability from the Census ACS, which returned data for a subset of state-years. States with above-median youth shares have an estimated ATT of +0.173 (SE = 0.467, $p = 0.71$); states with below-median youth shares have an ATT of −0.144 (SE = 0.314, $p = 0.65$). Neither estimate is significant. The small subsamples (due to ACS data gaps) substantially reduce statistical power.

E. Additional Figures and Tables