doi:10.1017/S0003055424000108

Navigating Potential Pitfalls in Difference-in-Differences Designs: Reconciling Conflicting Findings on Mass Shootings' Effect on Electoral Outcomes

HANS J. G. HASSELL *Florida State University, United States*
JOHN B. HOLBEIN *University of Virginia, United States*

Work on the electoral effects of gun violence in the U.S. relying on difference-in-differences designs has produced findings ranging from null to substantively large effects. However, as difference-in-difference designs, on which this research relies, have exploded in popularity, scholars have documented several methodological issues including potential violations of parallel-trends and unaccounted for treatment effect heterogeneity. These pitfalls (and their solutions) have not been fully explored in political science. We apply these advancements to the unresolved debate on gun violence's effects on U.S. electoral outcomes. We show that studies finding a large positive effect of gun violence on Democratic vote shares are a product of a failure to properly specify difference-in-differences models when underlying assumptions are unlikely to hold. Once these biases are corrected, shootings show little evidence of sparking large electoral change. Our work clarifies an unresolved debate and provides a cautionary guide for scholars currently employing difference-in-differences designs.

Gun violence in the United States has a devastating impact (e.g., Barney and Schaffner 2019; Hartman and Newman 2019; Marsh 2022; Rogowski and Tucker 2019; Rossin-Slater et al. 2020). Yet, despite repeated tragedies and public support for policies to reduce gun violence, policy response has been tepid (Goss 2010; Luca, Malhotra, and Poliquin 2020). This presents an unsolved puzzle. Why do salient mass shootings and a supportive public fail to instigate policy change? To solve this puzzle, scholars have examined whether mass shootings change electoral incentives and whether a lack of policy changes occurs *in spite of* or (perhaps) *because of* a lack of electoral pressure. In estimating the effects of these tragic shootings on elections, scholars have relied on panel data and difference-in-differences designs that exploit variation in shootings' timing and location. Yet, despite using the same data sources, previous work reaches starkly different conclusions, with some finding mass shootings have strong electoral effects (García-Montoya, Arjona, and Lacombe 2022; Yousaf 2021) and others finding null effects (Hassell, Holbein, and Baldwin 2020).

We show these conflicting findings come from the failure of some work to account for potential violations of the (essential) parallel trends assumption. Previous work documenting large effects of gun violence on electoral outcomes concludes so erroneously because shootings are more likely to happen in areas trending Democratic before shootings happened, whereas areas where shootings have not occurred were trending Republican.[1] (Note: if all the reader wants is a clear explanation of why research that does not account for parallel trends violations diverges in its conclusions from research that does account for trends, see Figures 2 [showing different trends for counties with and without shootings] and 4 [showing counties with shootings trending Democratic *even before* shootings]).

Models accounting for violations of parallel trends provide no evidence that mass shootings cause large electoral change in the United States, and while chances of much smaller positive *or negative* effects cannot be entirely eliminated, almost all of these estimates are not statistically significant and are highly dependent on specific model specifications (see Figure 11). Sensitivity analyses embracing uncertainty around exact departures from parallel trends show these results are *highly* sensitive to *minimal* reasonable departures from parallel trends. Hence, the preponderance of evidence does not support conclusions that mass shootings have any large positive effect on Democratic vote shares.

These results are consistent whether we look at all mass shootings, school shootings, or just "rampage-style" school shootings.[2] (Note: if all readers want out

Hans J. G. Hassell, Professor, Department of Political Science, Florida State University, United States, hans.hassell@fsu.edu.
Corresponding author: John B. Holbein, Associate Professor of Public Policy, Politics, and Education, Frank Batten School of Leadership and Public Policy, University of Virginia, United States, holbein@virginia.edu.

Received: December 09, 2021; revised: May 05, 2022; accepted: January 25, 2024.

[1] Likely because mass shootings have disproportionately occurred in recent years and in growing population areas (Musu-Gillette et al. 2018; U.S. Government Accountability Office 2020) coinciding with political realignments where these areas have become more Democratic (DeSilver 2016).

[2] While stating "[their] findings hold when [they] replicate [Hassell, Holbein, and Baldwin 2020; hereafter HHB] models with [their]

of this article is a clear *estimate of the effects* [or more appropriately the lack thereof] of gun violence on Democratic vote shares, see Figures 5 [showing nonsignificant effects after controlling for unit time trends], 6 [showing nonsignificant effects after controlling for unit time trends in event-study designs], and 11 [showing the distribution of effects around zero].)

Resolving discrepancies in these published findings also provides an opportunity to illustrate the critical importance of navigating pitfalls in difference-in-differences designs. The difference-in-differences design has recently proliferated, partially because of its simplicity and modest data requirements coinciding with a broader interest in causal inference and "credible" estimates which it provides (Angrist and Pischke 2010). This has prompted a growing methodological literature covering the potential and pitfalls of this design (e.g., De Chaisemartin and d'Haultfoeuille 2020; Kahn-Lang and Lang 2020; Roth et al. 2022).

Although we strive first to answer the question of whether mass shootings affect election outcomes in the United States, we are also interested in narrowing the gap between theory and application of the difference-in-differences designs in political science. We do so by (1) outlining potential biases arising from (a) violations of parallel trends and (b) treatment effect heterogeneity; (2) highlighting the importance of researcher decisions related to specifying difference-in-differences models (e.g., treatment coding, the use of time trends and, if so, their functional form, and how [or at what level] to adjust standard errors); and (3) implementing them in an applied example.

This article provides a guide in the application of the difference-in-differences design and provides an answer to an important unresolved debate, shedding light on the political economy of gun violence in the United States and contributing to our understanding of what events spark electoral accountability.

DIFFERENCE-IN-DIFFERENCES AND THE TWO-WAY FIXED EFFECTS ESTIMATOR

We think it is important to first explain the predominant difference-in-differences approach and the logic behind the pitfalls that exist. Difference-in-differences designs routinely rely on *two-way fixed effects estimator* (TWFE). With TWFE, the outcome of interest is regressed on time and unit (often geographic) fixed effects, along with the treatment status. The TWFE controls for factors remaining constant within years (e.g., nationwide economic conditions) and factors varying across spaces (e.g., stable local culture).

Original difference-in-differences designs used these identification strategies in largely exogenous interventions implemented in a single time period. This design constitutes a two-group (treated and not treated) and two-period (pre and post) design and "[the difference-in-differences estimator] is equal to the treatment coefficient in a TWFE regression with group and period fixed effects" (De Chaisemartin and D'Haultfoeuille 2023, C3).

Importantly, this design rests on the parallel trends assumption or the assumption that without treatment "that the average outcome among the treated and comparison populations would have followed 'parallel trends' in the absence of treatment."[3] As discussed below, there are multiple ways to evaluate and adjust for violations of this crucial assumption.[4]

DIFFERENCES IN FINDINGS ON THE ELECTORAL EFFECTS OF MASS SHOOTINGS

In an article published at the *American Political Science Review* (*APSR*), Hassell, Holbein, and Baldwin (2020) (hereafter HHB) estimate the effect of school shootings on voter turnout and election outcomes at federal, state, and local levels. Using various modeling strategies, HHB find school shootings—regardless of the number of victims—have precisely estimated null effects on vote shares.[5] In contrast, in a later article in *APSR*, García-Montoya, Arjona, and Lacombe (2022) (hereafter GMAL) focus on the effect of "rampage-style" school shootings. Using a TWFE, they emphasize (in the abstract and throughout the manuscript) that these shootings increase Democratic vote share by around 5 percentage points in the local community.[6] In work published in the *Journal of the European Economic Association*, Yousaf (2021) uses a TWFE showing all mass shootings—not restricted to school shootings—decrease Republican presidential vote share by 2–6 percentage points locally.[7]

data" (17) which include county-specific time trends, García-Montoya, Arjona, and Lacombe 2022 (GMAL) do not provide models with county-specific time trends in any of their models in their manuscript, appendix, or replication materials. Yousaf (2021) also does not include models with time trends. None, including HHB, apply more recent advances addressing potential parallel-trends violations.

[3] Roth et al. (2022, 2219) also note a related assumption requiring "the treatment has no causal effect before its implementation (no anticipation)."

[4] Tests of parallel trends are a part of a broader form of falsification testing (e.g., Keele 2015; Keele and Minozzi 2013), wherein one provides evidence against the validity of an identification strategy (acknowledging that evidence for that identification strategy is not possible). Even without appearing to fail parallel trends assumptions, issues can arise, for example, with compound treatments (Keele and Minozzi 2013). Moreover, parallel-trends tests can be underpowered (Bilinski and Hatfield 2018; Rambachan and Roth 2021).

[5] HHB also use regression discontinuity in time to assess shootings' effects on voter registration (also a null result).

[6] Depending on the sample one uses—either the full pool of observations or only ones where covariates are available—GMAL's naive TWFEs are 5.5 and 4.5 percentage points, respectively ($p < 0.001$ for both). GMAL run various specifications, which range in terms of the size of effects they document; being somewhere between 2 and 5 percentage points (see their Figure 4).

[7] Unlike HHB, neither GMAL nor Yousaf examine midterms or state and local races.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

Going one step further, there are model specifications that are similar to those that GMAL and Yousaf run on their data and are theoretically justified but that are not run in these original papers. Some of these plausible alternative specifications (albeit, these are not explored by GMAL or Yousaf) suggest effects as large as ≈ 8.7 percentage points the election following a mass shooting. In addition, some event-study models using TWFE specifications (albeit, these are not explored by GMAL or Yousaf) suggest effects as large as 13 percentage points a full 28 years after a mass shooting.

Ultimately, GMAL's and Yousaf's conclusions differ with HHB's—with the former suggesting large statistically detectable meaningful effects of mass shootings in partisan vote shares in local communities. Our work shows results from TWFE models suggesting these large effects are not robust because they fail to fully account for violations of the critical parallel trends assumption.

DATA/METHODS

All papers in this literature use a common dataset—Dave Leip's Atlas of U.S. Elections—which reports county-level vote shares.[8] We focus our examination on key differences in previous work examining the electoral effects of gun violence.[9] In estimating difference-in-differences designs, one decision in the hands of the researcher is what counts as treatment. In our case, each study uses slightly different shootings as treatments (see Supplementary Table S1). However, despite previous claims (GMAL, 821–3), differences in data choices and coding are ultimately not what drives divergent findings.

Below, we are interested not only in the statistical significance of the effects but also in their magnitude. While adjudicating effect size is always somewhat in the eye of the beholder, we use several tools to quantify the size of our observed effects. First, we benchmark estimates to other similar geographic-based treatments. Second, we use equivalence testing to see what effects we are able to rule out (Hartman and Hidalgo 2018). Finally, though not fully capturing the scope of effects, we also note when effects are not statistically significant.

TWFE ESTIMATOR IN OUR EMPIRICAL CASE

As detailed previously, the most common approach to estimating difference-in-differences effects when treatment varies over time and space is the TWFE. This approach is followed by GMAL and Yousaf, who emphasize results using county and year fixed effects.[10] We replicate this approach with the data provided by HHB although this is not their primary estimator.[11] The TWFE is specified in Equation 1, where Y_{ct} represents the Democratic vote share in a county (c) and election period (t), ϕ_c represents a county fixed effect, λ_t represents a year fixed effect, ϵ_{ct} represents the error term, and D_{ct} denotes the treatment (i.e., whether a county (c) in a given election period (t) is exposed to a mass shooting). β is the effect mass shootings on Democratic vote share:[12]

$$Y_{ct} = \phi_c + \lambda_t + \beta * D_{ct} + \epsilon_{ct}. \tag{1}$$

Another researcher decision point is how to code the exact nature of the treatment.[13] One possibility is to code units exposed to treatment in a given period as treated and all other observations—pre- and posttreatment in eventually treated units and never-treated units—as the control group. In the mass shooting example, this approach codes all counties with a mass shooting in

[8] HHB focus on school shootings occurring between 2006 and 2014 and, in robustness checks, between 2000 and 2018, GMAL focus on "rampage-style" shootings between 1980 and 2016, and Yousaf uses the FBI's definition of a mass shooting "leading to four or more deaths at one location" with data from 2000 to 2016.

[9] We are grateful to all the authors because we successfully replicated all reported findings using their code. As discussed below, our finding that mass shooting effects are insubstantial when one fully accounts for parallel trends is also corroborated by Marsh (2022). While Marsh does provide some evidence that mass shootings close to an election have a slight positive effect on turnout (see Marsh 2022, Figure 1), HHB show that the effects of school shootings close to an election on turnout are highly sensitive to model specification (see HHB, Figure A7), a pattern also somewhat evident in Marsh's models (see Marsh 2022, Table E2 and E5). Importantly, then, given the lack of any substantive effect on turnout, any increase in Democratic vote share should come from persuasion, rather than mobilization, unless gun violence simultaneously demobilizes Republicans and mobilize Democrats at the exact same rates, which is highly unlikely. However, any persuasive effect would also likely show up in attitudinal shifts and previous research on the attitudinal effects of mass shootings has disagreed whether attitudinal effects are present and, if they are, whether these effects are polarizing or a uniform leftward shift (Barney and Schaffner 2019; Hartman and Newman 2019; Rogowski and Tucker 2019). An absence of an attitudinal shift does not alone undermine GMAL and Yousaf's results (attitudes are not behaviors) but it provides a theoretical reason to question effects on vote shares.

[10] These papers also include some time-varying controls, but the bulk of identifying assumptions come from county and year fixed effects. In some specifications, Yousaf compares successful shootings with non-successful shootings and in others includes flexible population time trends. GMAL, in some specifications, use neighboring counties as the control group, state fixed effects instead of county fixed effects, or decade fixed effects as opposed to year fixed effects.

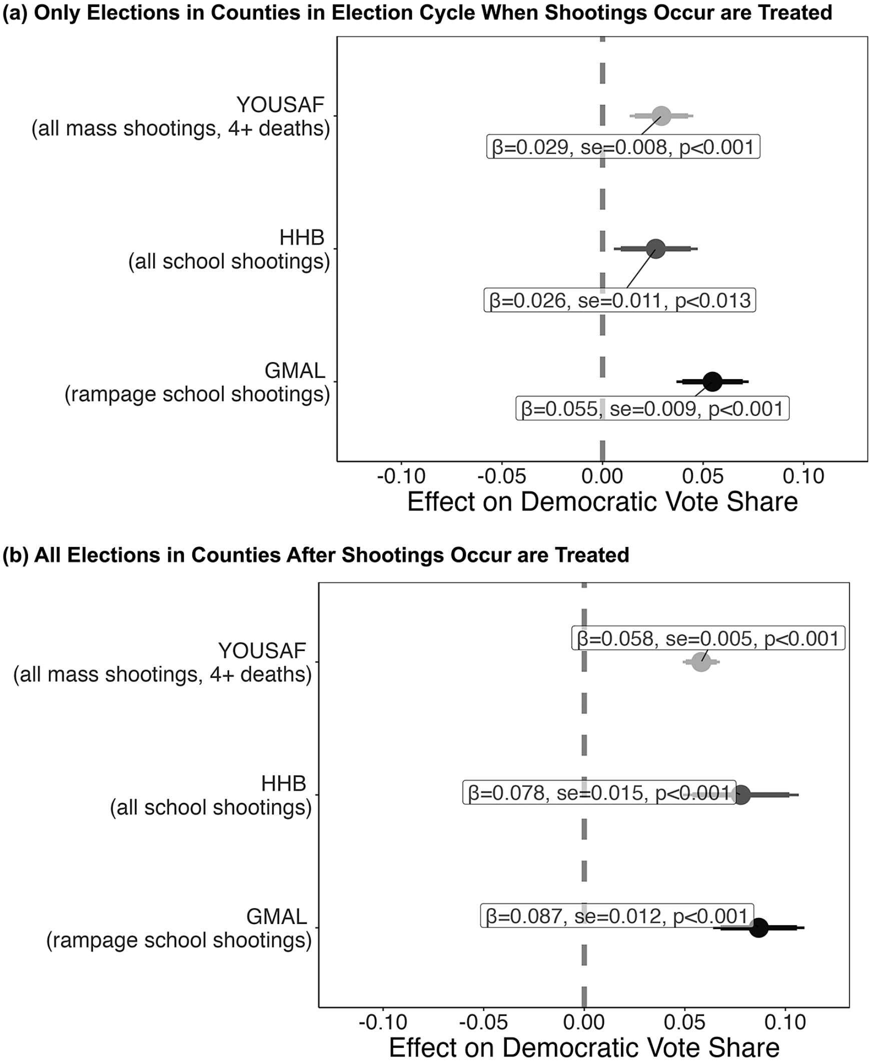
[11] HHB include unit-specific time trends.

[12] Another researcher decision (of less consequence here) is how to estimate standard errors (MacKinnon, Nielsen, and Webb 2023). We cluster standard errors at the treatment level (the county level). In general, we advise thoughtfulness in clustering standard errors (see Abadie et al. 2023; Cameron and Miller 2015).

[13] As HHB note, treatment does not need to be constrained to counties where shootings occur. HHB examine (and fail to find) effects in surrounding counties, as a function of the distance to a shooting, as a function of the severity of a shooting, and at the national level (with daily voter registration counts as the outcome).

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 1. Differences in Previous Studies' Estimated Effect of Mass Shootings on Election Outcomes Are Not Driven by Data Choices



Note: Estimates include county and year fixed effects (i.e., the TWFE estimator) with standard errors clustered at the county level and no covariates. The top panel shows effect estimates coding only the election immediately after a shooting occurs as having been treated; the bottom panel considers all post-shooting elections in counties with a shooting as treated. Coefficients, standard errors, and *p*-values are labeled for each coefficient. *Takeaway:* Naive TWFE estimators suggest mass shootings—regardless of the data/coding used—increase Democratic vote share in the county where shootings happens by 2.6–8.7 percentage points.

a given electoral cycle as treated, but county-level observations before and after that electoral cycle (along with those who never have a shooting) as untreated, allowing treatment in- and out-switchers. This approach assumes effects of mass shootings are constrained to the immediate electoral cycle. Alternatively, another approach is to code treatment, so all observations in treated units posttreatment are coded as treated. In our example, this approach codes all counties with mass shootings in an election cycle and following election cycles as treated, and all counties before—along with counties never having shootings—as untreated. This means there are no out-switchers. This approach allows mass shootings to have longer effects, changing the electoral environment both when they occur and afterward.[14]

The choice between the two approaches is a researcher decision that should be motivated by theory. Here, given a lack of strong expectations about mass shootings temporal effects, we use both approaches. (We complement these approaches with an event-study design described below, explicitly modeling effects in periods before and after shootings with lags and leads.)

[14] Supplementary Figures S13 and S14 provide visual illustrations of these two approaches (for a random sample of the observations) using the panelView package developed by Mou, Liu, and Yiqing (2022a).

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

TWFE Estimates of Shootings' Effect on Vote Shares

We start by showing that different conclusions across studies on the electoral effects of shootings are not driven by data choices. Figure 1 removes differences in methodological approaches in previous studies and shows the effects of TWFE models (GMAL's and Yousaf's approaches) using the data from all of the studies. Figure 1 also splits the results by treatment coding approaches outlined above.[15]

As shown in Figure 1, TWFE specifications consistently produce substantive positive statistically significant effects regardless of the time frame, treatment codings—be they "rampage-style" school shootings (GMAL), school shootings (HHB), or all mass shootings (Yousaf)—or how long treatments apply. With GMAL's data, we find—like GMAL did—that "rampage-style" shootings correlate with increases in Democratic vote share. This estimate (while varying by specifications) is 5.5 percentage points in the TWFE estimator for the first treatment coding (i.e., panel a) and is 8.7 percentage points in the second treatment coding (i.e., panel b). Both estimates are highly significant ($p < 0.01$). The effects for HHB and Yousaf are very similar. Simply, when using the same model choices, the effects of shootings of different types are consistently sizeable. In other words, previous differences in conclusions across studies of the electoral effects of gun violence are not due to choices about which shootings count as treatment. In short, TWFE estimates, regardless of the coding of shootings, indicate significant and substantively meaningful positive effects of shootings on Democratic vote share.

We pause to discuss effect magnitude. Upwards of an 8.7 percentage point shift in Democratic vote share is large—as are many of the other estimates. As GMAL note, these effects represent "a remarkable shift in an age of partisan polarization and close presidential elections" (GMAL, 809). We see how large these effects are by benchmarking them to other studies using county-level vote shares and difference-in-differences designs. For example, Sides, Vavreck, and Warshaw (2022, 709) estimate a six-standard-deviation shift in relative television advertising produces a 0.5-point change in two-party vote share. Hence, if we believe these results travel, GMAL's simple TWFE estimates indicate one school shooting has an effect on Democratic vote share equivalent to a shift of approximately 66–104 standard deviations in relative advertising. Using an economic comparison —the most common of retrospective voting treatments—Healy and Lenz (2017, 1423) show a "1 percentage point increase in mortgage delinquencies increases Democratic vote share by 0.33 percentage points." Thus, the effect of "rampage-style" shootings is roughly the equivalent to a 16.7–26.4 percentage point increase in mortgage delinquencies; or moving from a world with no delinquencies to one where about one-fifth of residents are at risk of losing their homes.

In short, TWFE models suggests gun violence—regardless of how shootings are coded—fundamentally reshapes electoral results in the local communities in which they occur. Is this sizable relationship causal and robust? Recent methodological developments provide us a guide to answer this question.

ADDRESSING ISSUES WITH TWO-WAY FIXED EFFECTS ESTIMATORS

Recent research has shown that simple TWFE models can be problematic for important reasons, including:

1. violations of the parallel trends assumption (e.g., Freyaldenhoven et al. 2021; Liu, Wang, and Xu 2024; Rambachan and Roth 2021) and
2. mistaken inferences derived from heterogeneity in treatment effects (e.g., Goodman-Bacon 2021; Sun and Abraham 2021).

We discuss these issues in order and apply solutions articulated in the literature using the example of mass shootings' seeming effects on Democratic vote share.

ASSESSING AND ADDRESSING PARALLEL-TRENDS VIOLATIONS

A core assumption to the difference-in-differences design is the parallel trends assumption. The parallel trends assumption asserts outcomes of interest from pre- to posttreatment would have moved in parallel across treated and untreated groups if not for treatment. If parallel-trends assumptions are violated, estimated effects are biased. There are several ways to assess the potential for differential pretreatment trends.

Checking for Visual Evidence of Differential Pretreatment Trends

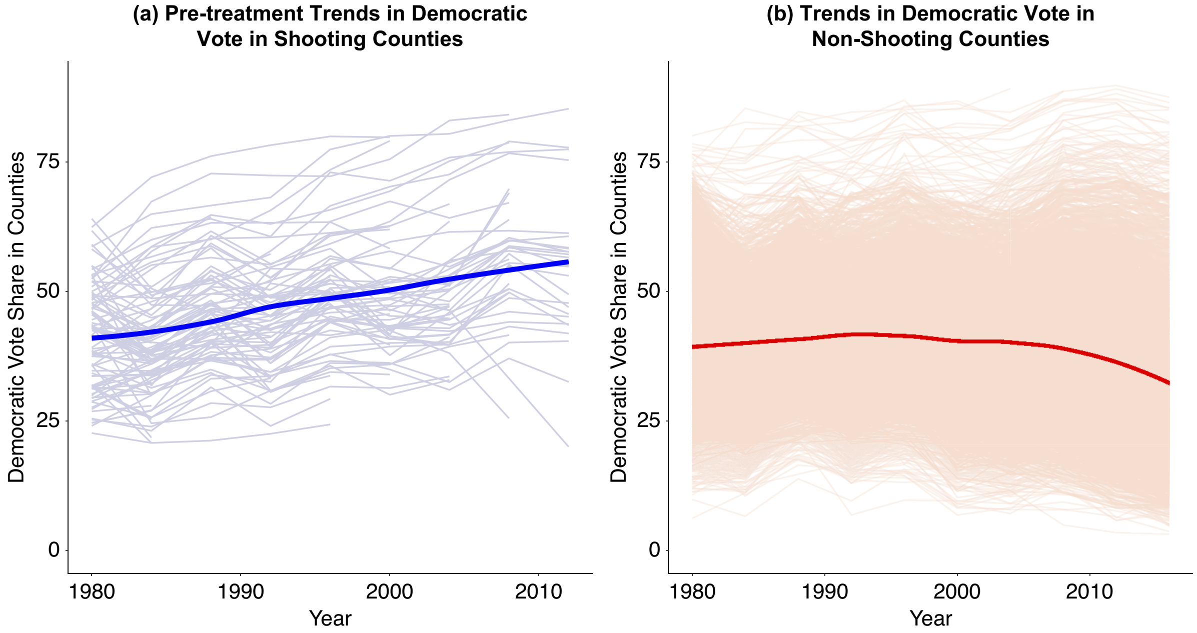
Because of the fundamental problem of causal inference, we do not observe counter-factual worlds where the treated and untreated groups are exposed to opposite conditions. Hence, no singular test can prove parallel trends is satisfied; however, treated and untreated units not moving together before treatment exposure, indicates potential issues (De Chaisemartin and D'Haultfoeuille 2023; Marcus and Sant'Anna 2021).[16]

[15] Table R1 in the Dataverse Files for this project provides estimates for Figure 1 (see Hassell and Holbein 2024).

[16] Marcus and Sant'Anna (2021) note whether pretreatment tests validate parallel trends assumptions "depends on the chosen [parallel] trends assumption." For other discussions of pre-trends tests, see Kahn-Lang and Lang (2020), Bilinski and Hatfield (2018), and Roth et al. (2022), and for potential relaxations of parallel trends

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 2. Trends in Presidential Vote in Counties with Mass Shootings Prior to Shootings, Compared to Trends in Counties without Shootings



Note: Pretreatment trends of Democratic vote share in counties where a shooting occurred (left panel) benchmarked to the trends in Democratic vote share found in counties where a shooting did not occur (right panel). Lighter lines show the patterns of individual counties; darker lines show the overall pattern for all counties. *Takeaway:* Counties that have shootings trended more Democratic *even before* the shootings occurred, whereas counties without a shooting trended slightly more Republican. Models that do not account for differential trends across counties will be biased.

An appropriate first step is to visually inspect patterns in aggregate-level data to see whether, prior to treatment, treatment and control areas are trending in different directions.[17]

Figure 2 examines differential pretreatment trends separating counties into two bins; panel (a) contains counties with a shooting excluding all post-treatment observations, and panel (b) contains all counties without shootings.[18,19] Figure 2 illuminates what TWFE models absorb and do not absorb. County fixed effects adjust for differences in Democratic vote share across counties. Year fixed effects account for differences across years. However, TWFE models *do not* account for the possibility counties' Democratic vote shares change at different rates over time, a significant problem in the context of school shootings.

As Figure 2 shows, shootings happen in communities —proceeding and unrelated to shootings themselves— trending more Democratic relative to other locales. This is likely because mass shootings occur disproportionately in growing populations and have increased over time (Musu-Gillette et al. 2018; U.S. Government Accountability Office 2020) at the same time American politics has realigned with these same more populated areas becoming more Democratic (DeSilver 2016). In general, researchers should take care when demographic/political changes predating treatment aligns with short-term treatment exposure.

This (coincidental) pretreatment trend separation becomes especially prevalent after 2004. This is particularly problematic because GMAL explicitly note shootings effects before 2004 are essentially null (or negative), but in 2004 the positive effects on Democratic vote share increase (GMAL, Figure 7). Figure 2 indicates this is the exact time when parallel trends assumptions become particularly tenuous. Two-way fixed effects models do not absorb these trends and, as such, are likely biased estimates.

Checking for Pretreatment Effects with the Model Specifications Used

While Figure 2 provides visual evidence of differences in pretreatment trends, it is not dispositive. Graphical representations can differ and changes in formatting can minimize or exacerbate the appearance of

assumptions, see Manski and Pepper (2018), Rambachan and Roth (2021), and Freyaldenhoven, Hansen, and Shapiro (2019).

[17] GMAL do this but fail to recognize potential violations of parallel trends in the bottom four figures for shootings after 2000 in GMAL's Supplementary Section A.4. Visual inspection is important but seldom sufficient.

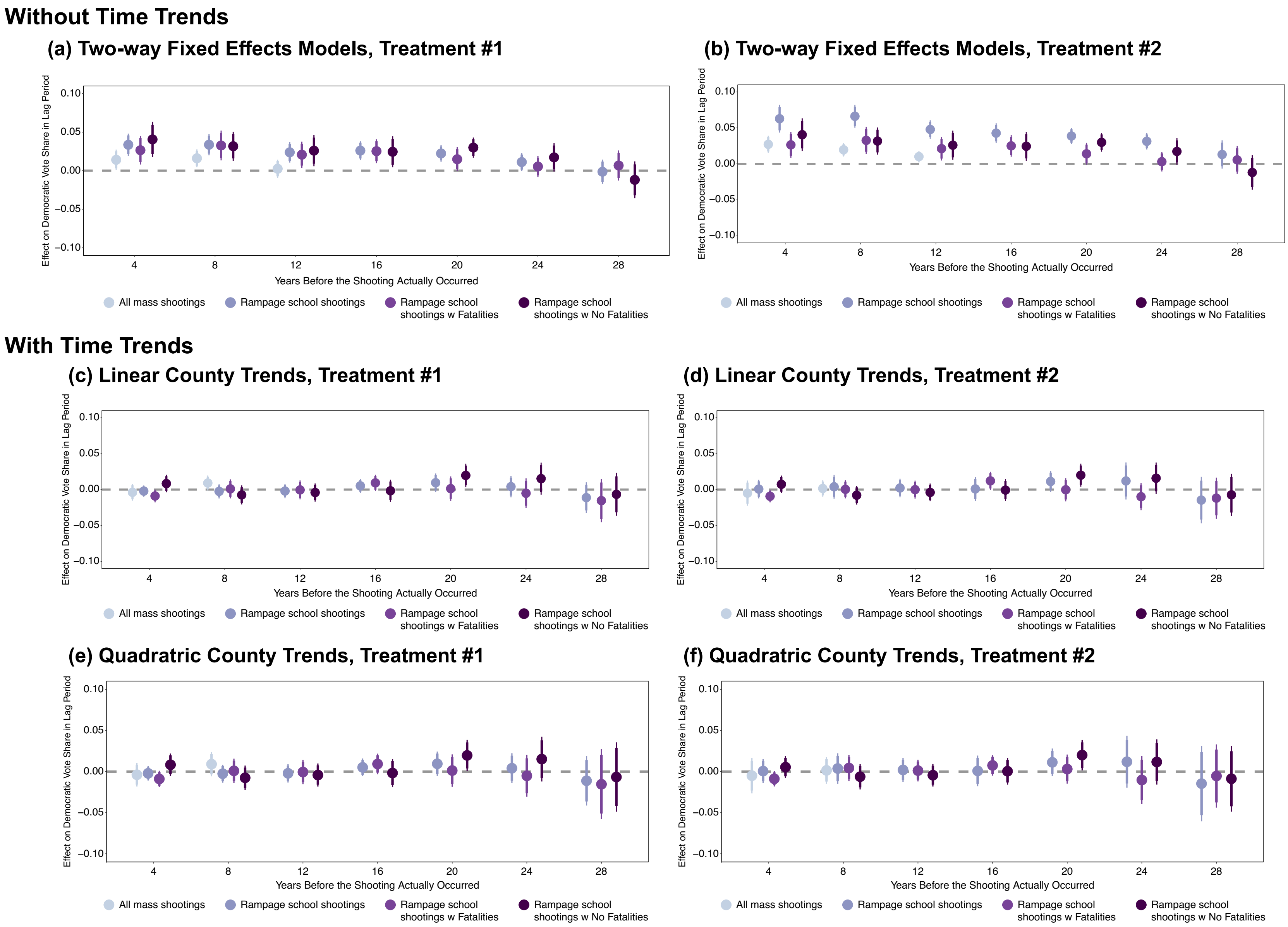
[18] Figure 2 uses GMAL's data. Supplementary Figure S10 shows Yousaf and HHB data.

[19] For the few counties with multiple shootings, we code the first shooting and treat all subsequent years as posttreatment.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 3. The Effect of Shootings on Election Outcomes Many Years Before



Without Time Trends

(a) Two-way Fixed Effects Models, Treatment #1

(b) Two-way Fixed Effects Models, Treatment #2

With Time Trends

(c) Linear County Trends, Treatment #1

(d) Linear County Trends, Treatment #2

(e) Quadratic County Trends, Treatment #1

(f) Quadratic County Trends, Treatment #2

Note: Effect of mass shootings on Democratic vote share in the years prior to when a shooting occurred. Treatment #1 is coded such that only elections with shooting are coded as treated; Treatment #2 is coded such that all elections after a shooting occurs in a county are coded as treated. All models' standard errors are clustered at the county level. *Takeaway:* TWFE estimators without time-trends indicate shootings may have an effect up to and including 20 years prior to when a shooting occurred.

differential trends leading researchers to draw different conclusions.[20] Hence, the next check to assess TWFE design validity which should be standard practice is whether this model suggests impacts *prior* to treatment (Grimmer et al. 2018). In our case, this placebo test is informative as shootings—something that people cannot precisely anticipate—should not affect vote shares prior to their occurrence. If there are effects, it suggests TWFE estimates are likely not causal (Angrist and Pischke 2008; Hansen and Bowers 2008).

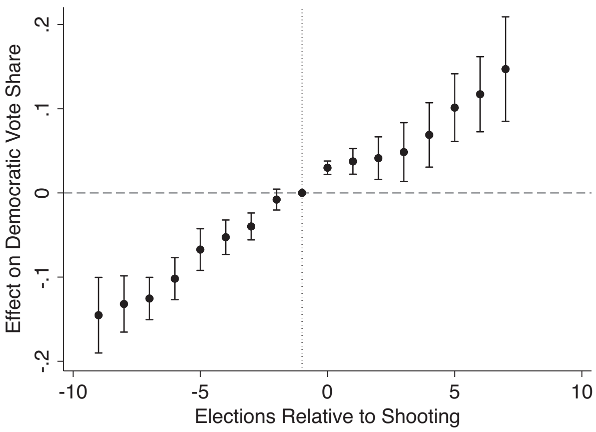
Specifically, we run the specification listed in Equation 2. Equation 2 is the same as Equation 1, except for the outcome variable. Instead of estimating shootings' effects (D_{ct}) in subsequent elections (Y_{ct}), we substitute a lagged outcome variable (Y_{ct-k}). Here, k corresponds to the number of lagged periods included. We include seven lagged periods in our models as the GMAL panel is sufficiently long. However, power considerations may influence the number of lags used. We recommend initially looking for effects in one period lag, then looking how far back one can estimate precise-enough specifications:[21]

$$Y_{ct-k} = \phi_c + \lambda_t + \beta * D_{ct} + \epsilon_{ct}. \tag{2}$$

Panels a and b (the top section) of Figure 3 (we discuss panels c–f later) show the TWFE models for various shootings on lagged measures of Democratic vote share and show there is substantial imbalance in lagged outcomes.[22,23] We start on the left of each panel, with the presidential election prior to the shooting and work up to seven presidential elections (28 years) before shootings occurred.[24] Effects vary by specification, but range between 2 and 7 percentage points, with most highly significant. This analysis indicates mass shootings have a significant and substantive effect on Democratic vote shares up to 20 years *prior* to a shooting. Simply, the TWFE does *not* recover balance prior to shootings, regardless of data used.

There is little reason—theoretically or empirically documented—to suspect shootings should have anticipatory effects on vote shares, given these events are unexpected where they occur. There is, however, the possibility pretreatment effects show up where there is not bias if treatment in one period (D_{ct}) is highly correlated with treatment in prior periods ($D_{c,t-k}$). In that case, the coefficient on D_{ct} may show an effect if there is an effect of $D_{c,t-k}$ on $Y_{c,t-k}$. In such cases, pretreatment effects could emerge even without errors in the research design. Therefore, we recommend also modeling effects of lagged and leaded treatment using an event-study design.

FIGURE 4. Event-Study Estimates Show that TWFE Fails to Account for Pretreatment Trends



Note: Event-study estimates with county and year fixed effects (GMAL's data). Baseline election year is shown with a gray dotted line. Following prior work, we bin our extreme points (Baker, Larcker, and Wang 2022; Schmidheiny and Siegloch 2019). *Takeaway:* Counties that have shootings trended more Democratic *even before* the shootings occurred. The increase that occurs after a shooting is entirely consistent with a general trend toward more Democratic election outcomes. Models that do not account for differential trends across counties will be biased.

Checking for Pretreatment Trends with Event-Study Designs

An event-study design traces effects before and after treatment and provides another way to see pretreatment imbalances (Armitage 1995; Binder 1998). An event-study is an increasingly common difference-in-differences model, given its less restrictive and more transparent modeling assumptions, but still relatively rare in political science. An event-study (usually) uses TWFE but also includes lagged and lead treatment variables as shown in Equation 3 below. We list treatment in a given county (c) and year (t), lagged or leaded by the corresponding periods since treatment (k). For simplicity, Equation 3 shows the event-study model for only one pretreatment period (k–2), the period when treatment occurs (k), and one period after treatment occurs ($k+1$). The baseline is the period before treatment occurs (k–1) (Armitage 1995; Binder 1998):

$$Y_{ct} = \phi_c + \lambda_t + \beta_{-2} * D_{ct,k-2} + \beta_0 * D_{ct,k} + \beta_1 * D_{ct,k+1} + \epsilon_{ct}. \tag{3}$$

Figure 4 shows nine preelection treatments and eight posttreatment periods included using the GMAL data. The right of Figure 4 (right of the gray vertical line) shows an immediate, significant, and substantive jump in Democratic vote share the election year following a shooting and grows after the event occurs, having an

[20] See, e.g., GMAL's Supplementary Section A.4 compared to Figure 2 presented here.

[21] Tests for imbalance are context-specific and scholars should consider the necessary precision of an imbalance using equivalence testing (Hartman and Hidalgo 2018).

[22] Supplementary Tables S13 and S14 provide the table version of panels a and b.

[23] HHB control for pretreatment trends, but TWFE estimate with their data also produces 2.2 percentage point increases in Democratic vote share 4 years prior ($\beta = 0.022$, $p < 0.073$).

[24] Yousaf's time frame only permits examining three Presidential election cycles prior.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

increasingly larger effect on Democratic vote share (10–13 points), and allows us to rule out smaller effects using equivalence testing. While not completely impossible, the long-lasting and growing effect remains theoretically unexplained.

However, by looking to the left of the baseline period (left of the gray dotted vertical line), Figure 4 also illuminates the "effect" is unlikely to be causal. If the TWFE models were causal, these coefficients *should not* be significantly/substantively distinct from zero. Instead, Figure 4 shows that relative to one election prior to a shooting, prior election years see less Democratic support and this underperformance increases further back in time. Simply, vote shares trend more Democratic before shootings in counties where shootings occur (relative to counties without shootings). Elections after a shooting are just a continuation—indeed, the points represent an almost perfect linear function—further evidence that TWFE estimators are biased in this case. We recommend parameterizing models as an event-study become standard in difference-in-differences applications.

Controlling for Any Differential Unit-Specific Time Trends

Facing potential parallel-trends violations, one potential remedy is adjusting for factors—observed or unobserved—leading to pretreatment imbalances. In this case, visual inspection (see Figure 2) reveals treated and untreated units have different pre-trends. A solution is to include unit-specific time trends (Angrist and Pischke 2008; 2010; Wing, Simon, and Bello-Gomez 2018), making the identifying assumption deviation from county-year trends. Identification comes from sharp deviations from otherwise smooth unit-specific trends corresponding to the following equation:

$$Y_{ct} = \phi_c + \lambda_t + \gamma_c * t + \beta * D_{ct} + \epsilon_{ct}. \tag{4}$$

While Equation 4 includes linear county-specific time trends ($\gamma_c * t$), the functional form is a potentially influential decision in the hands of a researcher.[25] As a result, we run many model specifications—all taking slightly different tacts to adjusting for differential pretrends. The next section shows various other approaches, including methods recently developed by Liu, Wang, and Xu (2024), Freyaldenhoven et al. (2021), and Rambachan and Roth (2021). We recommend scholars run robustness checks across several parameterizations of the unit-specific trends, acknowledging higher-order unit-specific trends could face a bias-variance tradeoff, especially in smaller datasets.

Figure 3c–f shows the pretreatment effect models with linear and quadratic trends (adding $\gamma_c * t^2$ to Equation 4). For cubic and quartic county-specific trends, see Supplementary Figures S9 and S10. In contrast to the TWFE (Figure 3a,b), all models with county-specific trends are balanced pre-treatment. Importantly, these null effects (especially in proximate periods) are not driven by standard error inflation ruling out even modest pretreatment differences using equivalence testing (effects outside the –0.99 to 0.57 percentage point range).

Figure 5 shows estimates for models including linear and quadratic county-specific time trends on the posttreatment outcomes with only counties with shootings in that year coded as treated.[26] As Figure 5 shows, once we make this necessary adjustment to correct for the violation of the parallel trends assumption, the effects of mass shootings of all types attenuate heavily. All effect estimates are *much* smaller and virtually all are not significant.

Specifically, Figure 5a,b—the estimates for "rampage-style" school shootings (i.e., GMAL's treatment)—indicates a 0.7 percentage point increase in Democratic vote share, an effect that is not statistically significant and 7.9 times (i.e., 790%) smaller than the original TWFE effectively ruling out effects as large as the TWFE with a high degree of confidence.[27,28] We can effectively rule out effects of "rampage-style" shootings (GMAL's treatment) larger than 1.5 percentage points and smaller than –0.1 percentage points, effects of all mass shootings (Yousaf's treatment) larger than 1.2 percentage points and smaller than −1.6 percentage points.[29] Effects of all school shootings (HHB's treatment) are also much smaller and not significant. Simply, there is no evidence of *large* effects documented in previous work finding an effect, and no consistent evidence for positive (or negative) effects statistically distinguishable from zero regardless of the data used.[30] While statistically significant effects infrequently show up in Figure 5, they are not robust. Notably, if we add higher-order polynomials—as in Supplementary Figure S9—no effects are significant. This—along with further checks below—emphasizes the *lack* of support for the conclusion that shootings have significant, systematic, or large effects on Democratic vote shares.[31]

[25] In another approach, we change the dependent variable to the change in Democratic vote share from the election before shootings occurred to the election in which counties are actually treated thus skirting Nickell bias arising from models with lags and fixed effects (Beck, Katz, and Mignozzetti 2014).

[26] Supplementary Tables S19–S22 provide model estimates for these figures and Supplementary Figure S14 shows results when coding counties with shootings as not reverting to the control afterward.

[27] Supplementary Table S11 shows including covariates with a linear time trends makes effects even smaller—0.2 percentage points ($p = 0.596$; 95% CI: [–0.7, 1.2]).

[28] Using the second treatment coding, the effect is 7.3 times smaller.

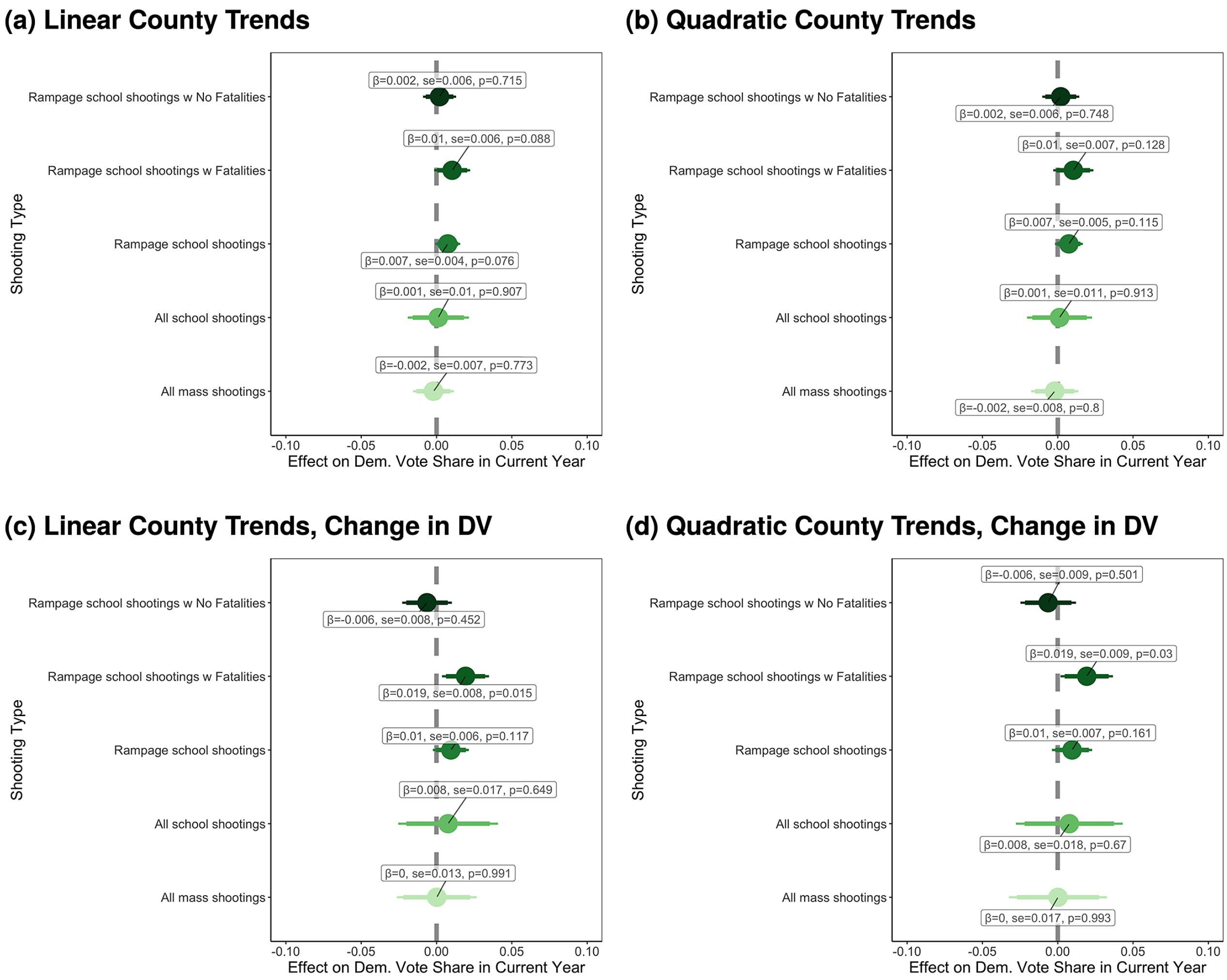
[29] As shown in Supplementary Figure S14, Yousaf's effects are not significant in three-fourths of the models run with the second treatment coding.

[30] Adding time trends might artificially inflate standard errors to unpalatable levels. However, confidence intervals remain small in models with linear time trends. They are slightly less precise with quadratic time trends, but are, to our eye, still quite tight with higher-order specifications shown in the Supplementary Material.

[31] Moreover, the null effects (once accounting for time trends) continue in robustness checks run by GMAL and Yousaf. Adding linear or quadratic time trends to Yousaf's original comparison of shootings versus failed shootings, reduces estimates ranging from 0.04 to 1.4 percentage points, with none close to statistical significance. Effects are also smaller and insignificant after adding linear

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 5. Effects of Mass Shootings on Elections after Absorbing County-Specific Trends



(a) Linear County Trends

(b) Quadratic County Trends

(c) Linear County Trends, Change in DV

(d) Quadratic County Trends, Change in DV

Note: Effect of mass shootings of various types once we account for differential trends. Within each panel, the first three estimates are using the GMAL coding of mass shootings and their data, the next comes from HHB, and the last comes from Yousaf. For cubic and quartic specifications, see Supplementary Figure S9. For effects where we code all post-shooting counties as being treated—not just counties and years with shootings—see Supplementary Figure S14. *Takeaway:* Once we account for differential trends across counties, the effects of mass shootings are all smaller and precisely estimated.

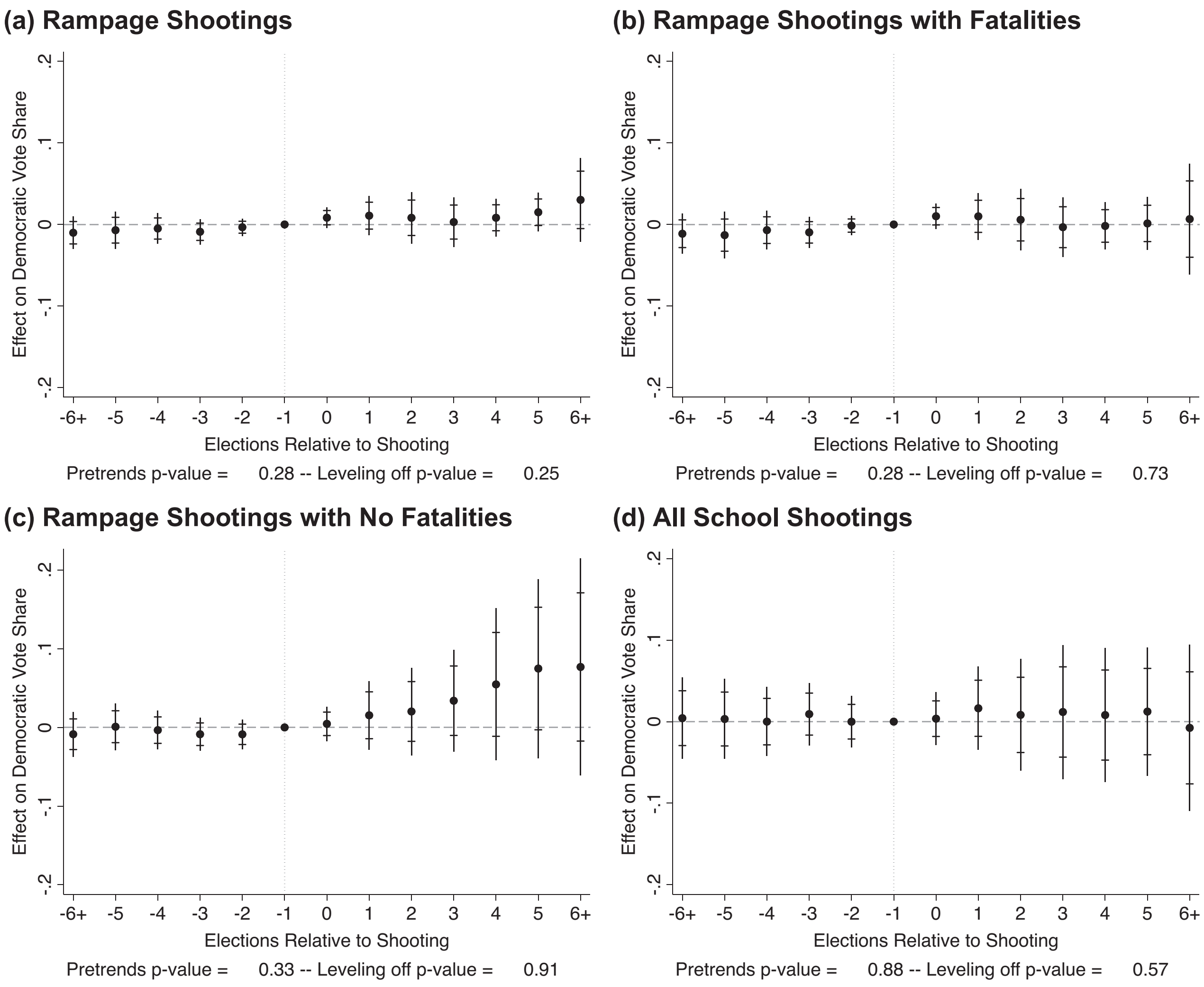
Using an event-study design that accounts for the differential pre-treatment trends by adding leads and lags of the treatment values bolsters conclusions as effects are even smaller and more precise. Figure 6 uses Freyaldenhoven et al. (2021) methods displaying event-study designs and accounting for pre-trends in event-study designs.[32,33] Figure 6 uses the same y-axis as Figure 4 for ease in comparison. Adding trends attenuates pre-treatment imbalances. So too, however, are any large post-treatment differences as the immediate effect of "rampage-style" school shootings is a mere 0.8 percentage point bump for Democrats. This effect is barely statistically significant using the linear trends models (p=0.048) but still allows us to confidently rule out large effects; we can rule out effects larger than 1.67 percentage points using equivalence testing, nowhere near the size of GMAL's 5 percentage point estimate emphasized throughout their text. Using the quadratic trends model, the effect is only marginally significant at the 10% level (p<0.066) and we can rule out effects larger than 1.71 percentage

or quadratic time trends to GMAL's original estimates using neighboring counties ($\beta = 0.5$ percentage points; $p < 0.61$ and $\beta = 0.5$ percentage points; $p < 0.6$, respectively). GMAL also run model with state and decade fixed effects (which do not absorb county-specific factors). Effects also attenuate dramatically if we add trends to these models.

[32] Supplementary Tables S23–S26 provide coefficients for these figures.

[33] This approach codes treatment only in the time period when the shooting occurred.

FIGURE 6. Event-Study Estimates of Shootings after Absorbing County-Specific Trends



(a) Rampage Shootings

(b) Rampage Shootings with Fatalities

(c) Rampage Shootings with No Fatalities

(d) All School Shootings

Note: Event-study estimates from the HHB and GMAL data with county and year fixed effects and county-specific quadratic time trends. These use the method developed by Freyaldenhoven et al. (2021) to account for pre-trends in event-study designs. Analysis executed using the xtevent and xteventplot commands in STATA (Freyaldenhoven et al. 2022). These commands, as a default, plot both the standard confidence intervals and those developed by Olea, Luis, and Plagborg-Møller (2019), which were developed for contexts with dynamic effects. The figure uses the same *y*-axis as Figure 4 for ease in comparing across the two. *Takeaway:* Once time trends are taken into account, the effect of shootings attenuates considerably.

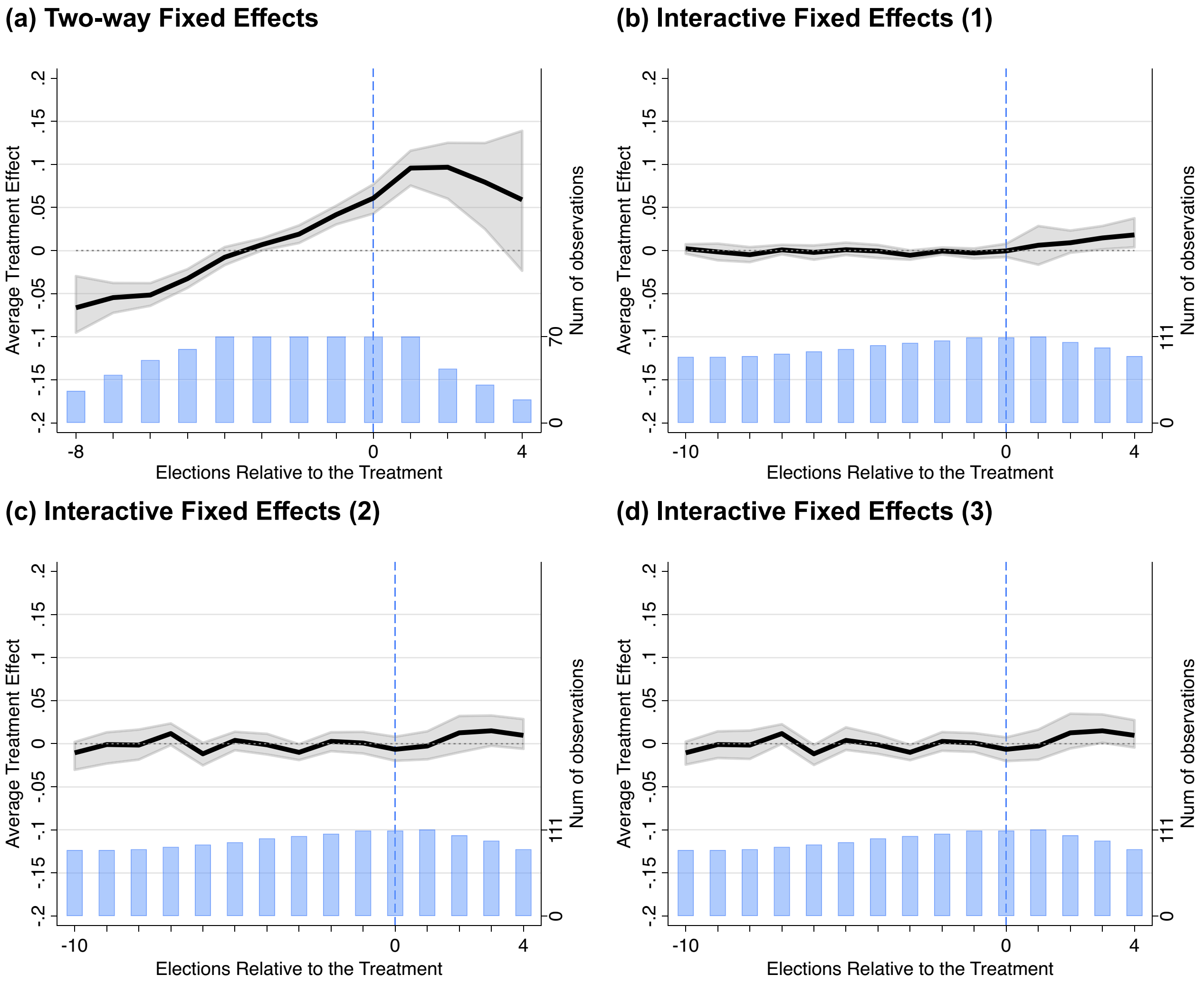
points.[34] These effects appear to be the upper bound produced from this method.

If we use Clarke and Tapia-Schythe's (2021) approach to estimating the event-study with trends and the corresponding eventdd command in STATA, we get *negative* effects on Democratic vote shares (albeit statistically indistinguishable from zero). With this slightly different approach, the effects in elections immediately following shootings are −0.13 percentage points ($p = 0.898$; 95% CI: [−2.05, 1.80]).[35] Other shooting codings are similarly very small and insignificant. Moreover, testing the robustness of these effects to other pretreatment periods—as the approach designed by Freyaldenhoven et al. (2021) allows—effects are even smaller and even less suggestive of an effect (see Supplementary Figures S2–S5). Benchmarked to the two-period lag trend, the estimate for elections immediately after a shooting using linear county trends is 0.47 percentage points ($p = 0.324$; 95% CI: [−0.5, 1.4]). Estimates for elections after a

[34] In the GMAL data, the evidence for effects shows up in "rampage-style" shootings with killings ($\beta = 1.0$ percentage points; $p = 0.064$; 95% CI: [−0.05, 2.1]) rather than "rampage-style" shootings without killings ($\beta = 0.45$ percentage points; $p = 0.550$; 95% CI: [−1.0, 1.9]).

[35] Quadratic county-specific trends are: −0.09 percentage points ($p = 0.927$; 95% CI: [−2.1, 1.90]).

FIGURE 7. Liu, Wang, and Xu (2024) Interactive Fixed Effects Counterfactual Estimator


(a) Two-way Fixed Effects

(b) Interactive Fixed Effects (1)

(c) Interactive Fixed Effects (2)

(d) Interactive Fixed Effects (3)

Note: The interactive fixed effects counterfactual estimator developed by Liu, Wang, and Xu (2024) using GMAL's data. Panel a shows the TWFE estimated by Liu, Wang, and Xu's (2024) FECT package; it is analogous to Figure 5, but their procedure estimates slight differences—for example, in the number of pre- and posttreatment periods. In panel b, the number of factors (*r*) is set to 3—that chosen by cross-validation and the degree of the polynomial is set to 4. In the bottom row, *r* is set to 1 in both panels and degree 2 in panel c and 4 in panel d. For other variations, see the Supplementary Material. *Takeaway:* The upward trend in the TWFE model (i.e., panel a) is indicative of violation of the parallel trends assumption. In the interactive fixed effects models, there is no evidence of the substantial effects shown in more simplistic model specifications that do not account for potential violations of the parallel trends assumption.

shooting for the quadratic county trends model is 0.5 percentage points ($p = 0.359$; 95% CI: [−0.5, 1.5]). Moreover, none of the large longer-term effects remain.

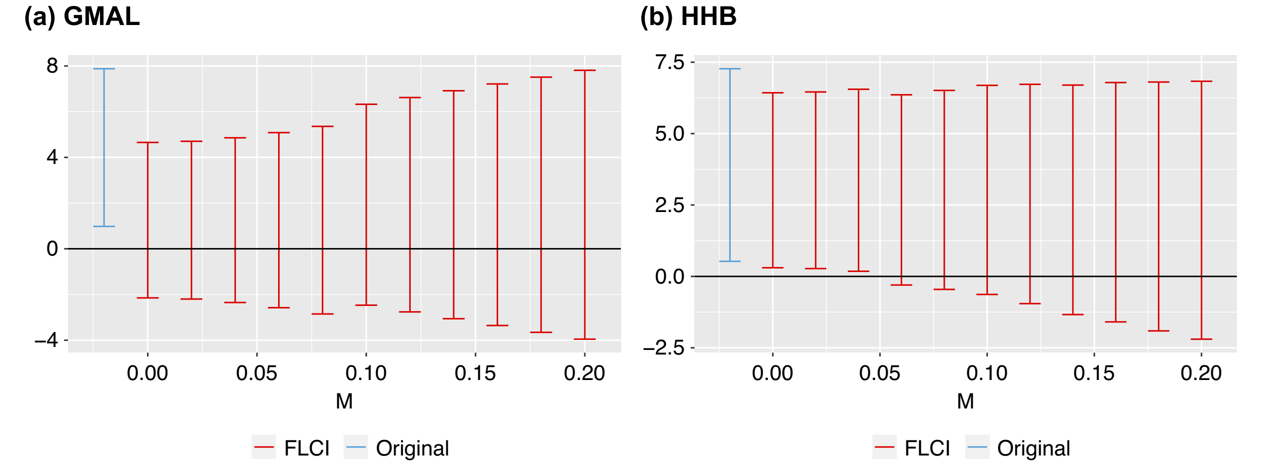
In short, in event-studies adjusting for unit-specific trends, there is little evidence for the sizeable effects previously suggested. In fact, there is *little evidence for any significant effect.* The occasional effect crossing the $p < 0.05$ threshold are not robust to reasonable model variations, such as the baseline comparison points one uses. Moreover, while some specifications cannot fully exclude some much smaller but non-negligible positive effects, most 95% confidence intervals also cannot rule out non-negligible *negative* effects.

Additional Checks Addressing Violations of Parallel Trends Assumptions

Including unit-specific time trends (see above) is not the only solution to parallel-trends violations; indeed, scholars may desire more flexible solutions. Recent advances have suggested many alternative solutions to potential parallel-trends violations or unobserved time-varying confounders. Unfortunately, these approaches have not yet been benchmarked to one another, let alone compared in different data contexts. The standard approach is to develop a new estimator and assert it applies in all contexts. We are currently

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 8. Implementing Rambachan and Roth's Sensitivity Analysis in the Shooting Example



Note: Results from the sensitivity analysis suggested by Rambachan and Roth (2021) using GMAL's data—that is, testing for effect sensitivity across $\triangle^{SD}(M)$. The models incorporate information from three elections prior to treatment and five post-treatment periods. *Takeaway:* The results show the effects of shootings on vote shares are highly sensitive and do not hold with even minor deviations from parallel trends.

unaware of work testing these methods head-to-head, let alone providing recommendations regarding which methods to apply in different contexts. As such, we recommend scholars run a variety of specifications, as we do below, identifying general patterns. We recommend that scholars implement, at minimum, checks suggested by Liu, Wang, and Xu (2024) and Rambachan and Roth (2021) outlined below.

Building on research exploring factor-augmented models for causal identification (e.g., Bai and Ng 2021; Xu 2023), Liu, Wang, and Xu (2024) develop procedures—including what they call the fixed effects counterfactual estimator, the interactive fixed effects counterfactual estimator, and the matrix completion estimator—to "estimate the average treatment effect on the treated by directly imputing counterfactual outcomes for treated observations" (1).[36] Using simulations, Liu, Wang, and Xu (2024) show that the interactive fixed effects counterfactual estimator provides more reliable causal estimates than conventional TWFE models when unobserved time-varying confounders exist. The interactive fixed effects counterfactual estimator can be applied with the package panelView, which is available in both Stata and R (Mou, Liu, and Yiqing 2022b).

Figure 7 applies Liu, Wang, and Xu's (2024) approach using the GMAL data.[37] Panel a shows the TWFE and panels b–d show interactive fixed effects counterfactual estimators. In the TWFE model, there are pretreatment imbalances and a general overall upward trend—as Figure 5 indicated previously. Again, this suggests the TWFE picks up on a general pro-Democratic trend in pre-treatment periods. However, after adjusting for pretreatment differences using Liu, Wang, and Xu's (2024) approach to address pretreatment imbalances, there is virtually no evidence shootings substantially or significantly affect vote shares in subsequent elections. Nor is this for lack of statistical power and we can rule out larger effects using equivalence testing. Moreover, any (much smaller) effects that appear intermittently are not robust to reasonable model variations in the realm of researcher decision-making.

Rambachan and Roth (2021) propose another solution using a sensitivity analysis approach for potential violations of the parallel trends assumption allowing researchers to avoid arbitrarily choosing a parametric model.[38] This approach is particularly useful as researchers often struggle to know the functional form of the underlying system.

This sensitivity analysis can be formalized in several ways. For example, researchers can see how robust their effects are to varying departures from differential trends evolving smoothly over time. This may be especially useful when concerned "about confounding from secular trends … evolv(ing) smoothly over time" (Rambachan and Roth 2021, 13)—as we are in this case. This sensitivity test is "done by bounding the extent to which the slope may change across consecutive periods" (Rambachan and Roth 2021, 12).[39] They call this the *SD* or "second derivative" or "second differences"

[36] This approach treats all units posttreatment as treated.

[37] See also Supplementary Tables S27–S30.

[38] This approach codes treatment only in the proximate time period.

[39] Under this approach, "the parameter M governs the amount by which the slope … can change between consecutive periods, and thus bounds the discrete analog of the second derivative" (Rambachan and Roth 2021, 13).

approach.[40] Using Rambachan and Roth's (2021) general approach, conclusions do not depend on arbitrary model specification choices. In essence, this approach "show[s] what causal conclusions can be drawn under various restrictions on the possible violations of the parallel-trends assumption" (Rambachan and Roth 2021, 1). This approach is implementable through the HonestDID package in *R* and *STATA* (Rambachan and Roth 2021; Rambachan, Roth, and Bravo 2021).

Rambachan and Roth (2021, 28) note "it is natural to report both the sensitivity of the researcher's causal conclusion to the choice of this parameter and the 'breakdown' parameter value at which particular hypotheses of interest can no longer be rejected." Figure 8 shows Rambachan and Roth's sensitivity approach using GMAL's and HHB's data. Figure 8 uses the *SD* approach, plotting robust confidence sets for the treatment effect in the mass shooting case for different values of the parameter M. The confidence sets show that the effect of mass shootings on Democratic vote share is only positive and significant in the coefficient on the far left—the original estimate not allowing for *any* violations of parallel trends—indicating effects of mass shootings are *highly* sensitive to any minor departures from parallel trends. The robust confidence intervals include zero when allowing for linear violations of parallel trends ($M = 0$), and become even wider allowing for nonlinear violations of parallel trends ($M > 0$). Such a low breakdown suggests any meaningful departure of the slope changing between consecutive periods, would cause the observed effects in GMAL's data to not be significant. Overall, these results indicate effects of shootings (using the GMAL data) are *highly* sensitive to even minor parallel-trends violations.

Summarizing Tests for Violations of Parallel Trends

Testing for potential violations of parallel trends is often not easy or straightforward because of the fundamental problem of causal inference—that is, the inability to observe what would have happened to treated groups without treatment. But even when considering only tests of pre-trends, there are challenges. For example, in choosing methods to address this core issue, one must consider the nature of the data available, the statistical power, and numerical degrees of freedom (Bilinski and Hatfield 2018; Roth 2022). We are *not* arguing every case employ unit-specific trends. What we *are* arguing is that all researchers should diagnose and address potential violations of this core assumption. *How* they do so—with the many tools at their disposal that we have outlined above—is less important than *that* they do so in a way that is transparent, thorough, and appropriate to the applied case.

In the case of mass shootings, this much is clear: once we make adjustments for clearly differential pre-trends, the evidence for any effect gets much more muddled than previous studies have suggested. Effects are substantively smaller than what simple TWFE models suggest. There is no clear evidence of durable lasting effects. Moreover, depending on the dataset one uses and the way treatment is coded, effects can be marginally positive, negative, or (approximately) zero, are very rarely significant, and highly sensitive to any meaningful departures from strictly parallel trends.

DIAGNOSING AND ADDRESSING TREATMENT EFFECT HETEROGENEITY/ REMOVING CONTAMINATED COMPARISONS

Recent research also indicates issues with variations in treatment with heterogeneity of treatment effects. For example, Goodman-Bacon (2021) shows any TWFE estimate with variations in treatment timing can be decomposed into an average of all possible 2 × 2 difference-in-differences estimates constructed from the panel data set weighted by group sizes and variance in treatment (Goodman-Bacon 2019). If there are time-varying treatment effects, they can produce biased estimates (Goodman-Bacon 2019, 3).[41] To diagnose this potential for bias, Goodman-Bacon (2019) allows for decomposing the 2 × 2 difference-in-differences estimates using the bacondecomp package in *R* and *STATA* (Goodman-Bacon 2021; Goodman-Bacon, Goldring, and Nichols 2019).[42] As stated in the package's *STATA* help file:

> [The decomposition] by default produces a graph for all comparisons and shows up to three types of two-group/two period comparisons, which differ by control group: (1) Timing groups, or groups whose treatment stated at different times can serve as each other's controls groups in two ways: those treated later serves as the control group for an earlier treatment group and those treated earlier

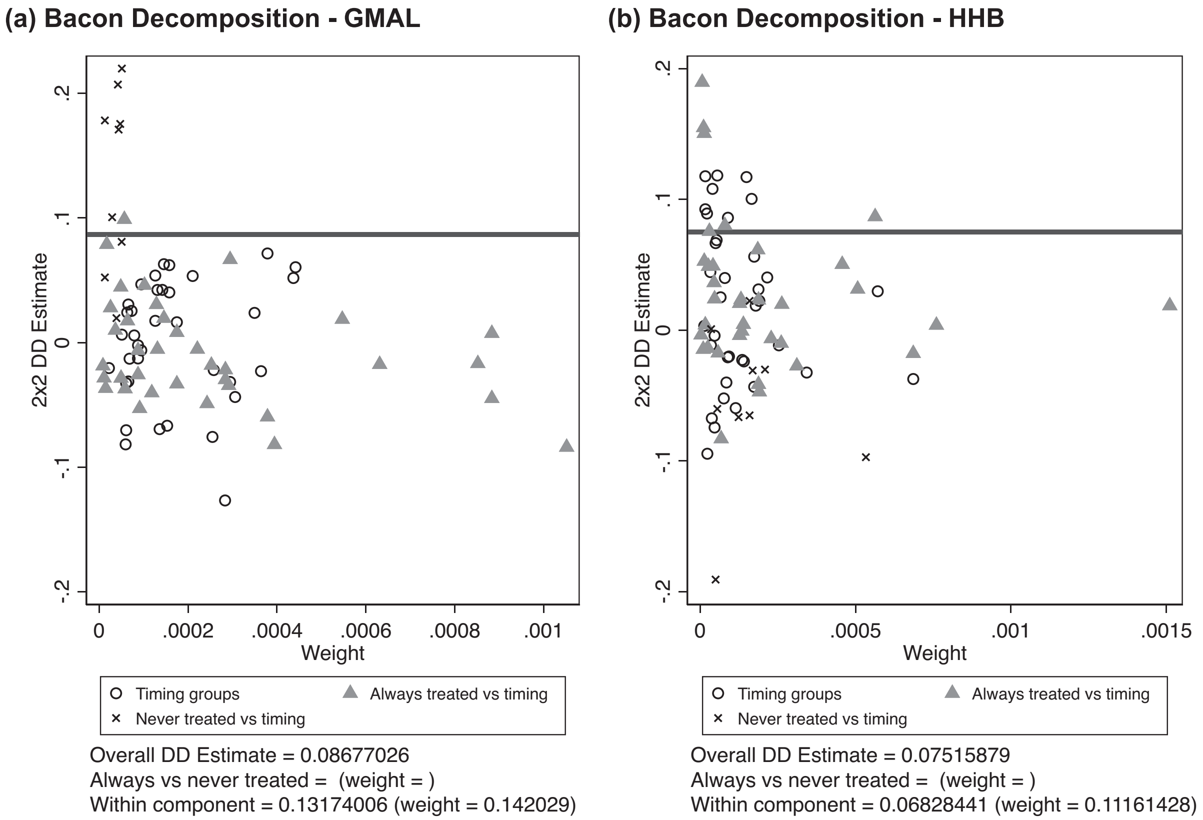
[40] In addition, researchers can see how robust effects are to (unobserved) posttreatment departures of parallel trends by benchmarking to (observed) maximum pretreatment violations of parallel trends. This is the *RM* or "relative magnitudes" approach. Researchers choose different values of $\bar{M}$, which measures how much of the maximum pretreatment violation of parallel trends would lead effects to include null effects in the confidence set. Rambachan and Roth argue this approach is reasonable—for example, if "possible violations of parallel-trends are driven by confounding … (shocks) of a similar magnitude to confounding … shocks in the pre-period" (Rambachan and Roth 2021, 12). Further, researchers can combine approaches in the *SDRM* condition. This approach "assume[s] … possible non-linearities in the post-treatment difference in trends are bounded by the observed non-linearities in the pre-treatment difference in trends" (Rambachan and Roth 2021, 13).

Under this approach, $\bar{M}$ is the parameter the research varies, allowing researchers to set "bounds [for] the maximum deviation from a linear trend in the post-treatment period by $\bar{M} \geq 0$ times the equivalent maximum in the pre-treatment period." *SDRM* is similar to *SD*, but "allows the magnitude of … possible non-linearity to explicitly depend on … observed pre-trends" (Rambachan and Roth 2021, 13).

[41] The rationale is explained in Goodman-Bacon (2021). See also Cunningham (2021, chap. 9).

[42] Here, we balance the panel and code posttreatment units as treated.

FIGURE 9. Illustration of the Goodman–Bacon Decomposition of the TWFE Models



Note: This figure shows the results from the Bacon decomposition for the TWFE models. The figure also shows all of the possible 2 × 2 difference-in-differences (DiD) estimate, with their weights for the ATE on the *x*-axis and the effect size on the *y*-axis. The horizontal line shows the overall DiD estimate.

serve as the control group for the later group; (2) Always treated, a group treated prior to the start of the analysis serves as the control group; and (3) Never treated, a group which never receives the treatment serves as the control group.

Figure 9 provides this decomposition.[43] As shown in both the GMAL and HHB data, the TWFE is a composite of 2 × 2's eliciting large negative and positive effects. Effects can be very different depending on the 2 × 2 's included in the estimate (as indicated by the spread of estimates across the *y*-axis). Second, the TWFE is a weighted composite highly influenced by several comparisons of always treated versus timing. However, many of the 2 × 2 estimates have similar weight—as noted by the cluster of estimates on the left side of the graph. Overall, to our eye, there appears to be no clear evidence our effects are driven by treatments of various types.

We think it is important to note that the mass shooting example is not an ideal application to show the value of Goodman–Bacon decomposition as (at present) the Goodman–Bacon decomposition only decomposes the TWFE and does not decompose more sophisticated models implemented accounting for parallel-trends violations. Nevertheless, this may not always be true in all contexts, and scholars should examine this decomposition to illuminate whether effects are driven by specific comparisons and if the bias Goodman-Bacon (2021) discusses is present.

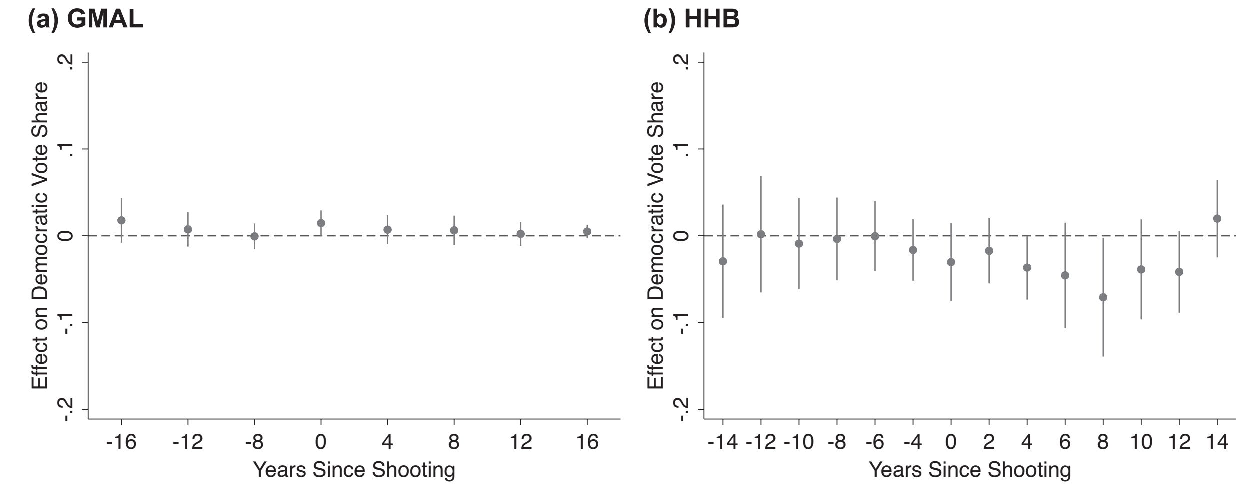
There are several proposed solutions to problems arising with heterogeneous treatment effects (Borusyak, Jaravel, and Spiess 2021; Callaway and Sant'Anna 2021; De Chaisemartin and d'Haultfoeuille 2020; de Chaisemartin, D'Haultfoeuille, and Guyonvarch 2019; Zhang 2022). Some allow the inclusion of additional covariates including unit-specific trends. One approach that does is Sun and Abraham (2021) implemented using the eventstudyinteract package in *STATA* and fixest package in *R* (Berge, Krantz, and McDermott 2022; Sun 2021).[44,45] This approach "estimates the shares of cohort as weights." In our case, implementing Sun and Abraham's solution with

[43] Weights for Goodman–Bacon decomposition are in Supplementary Tables S33 and S34. Supplementary Tables S8 and S9 show results using de Chaisemartin, D'Haultfoeuille, and Guyonvarch's (2019) alternative approach.

[44] With this approach, we make the panel strongly balanced and use the treatment in the current period coding.

[45] Results are robust to excluding never-treated observations.

FIGURE 10. Sun and Abraham (2021) Approach for Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects



Note: Results from the clean comparisons suggested by Sun and Abraham (2021) using the GMAL and HHB data. Models include quadratic county-specific time trends to address potential violations of the parallel trends assumption in the TWFE. *Takeaway:* Clean comparison effects with trends show no sign of a sizable and durable effect on Democratic vote shares shown in the TWFE nor in the simple event-study plot (see Figure 5).

a simple TWFE (with no time trends and thus not addressing parallel-trends violations) substantially reduces effect estimates (Figure 10). At first, these changes look modest with event-study estimates going from 3 percentage points ($p = 0.000$) in the naive models to 2.4 percentage points ($p = 0.001$) in the Sun and Abraham adjusted models in the first posttreatment period. However, in the second and following treatment periods, effects that were large (10–13 percentage points) heavily attenuate and even become negative (1.1 [$p = 0.19$], –0.4 [$p = 0.75$], –2.2 [$p = 0.12$], and –1.5 [$p = 0.18$] in posttreatment elections 1–4, respectively). Adding unit-specific trends to Sun and Abraham's estimator (panel a) makes effects even smaller. We go from 2.4 percentage points [$p = 0.001$] in the TWFE to roughly half (linear trends [$p = 0.031$]; quadratic [$p = 0.051$]; or cubic trends [$p = 0.085$]). (Even in the cubic model, the standard error remains modest in size —being 0.8 percentage points.) Moreover, the long-run effect of 10–13 percentage points is not present. None of the effects are present in the HHB data (panel b). This suggests that effect heterogeneity plays some role. Once adjusted for, long-term effects attenuate substantially and short-term effects become much smaller and flimsier to reasonable alternative specifications (e.g., the coding of treatment or the functional form of unit-specific trends).

More importantly, for reasons we outline below, it is unwise to overemphasize one model specification. Combining the evidence from all of the various approaches taking into account potential contamination from treatment effect heterogeneity, the best evidence suggests that in the case of shootings and electoral vote shares, (1) violations of parallel trends loom large, (2) effect heterogeneity may play a modest role, and (3) there is no sign of the sizable and durable effect, but perhaps a much smaller effect—and one that is *not* clearly distinct from zero across almost all reasonable model specifications (and is negative in some specifications).

OTHER POTENTIAL ISSUES RAISED IN THE LITERATURE ABOUT EXECUTING DIFFERENCE-IN-DIFFERENCES DESIGNS

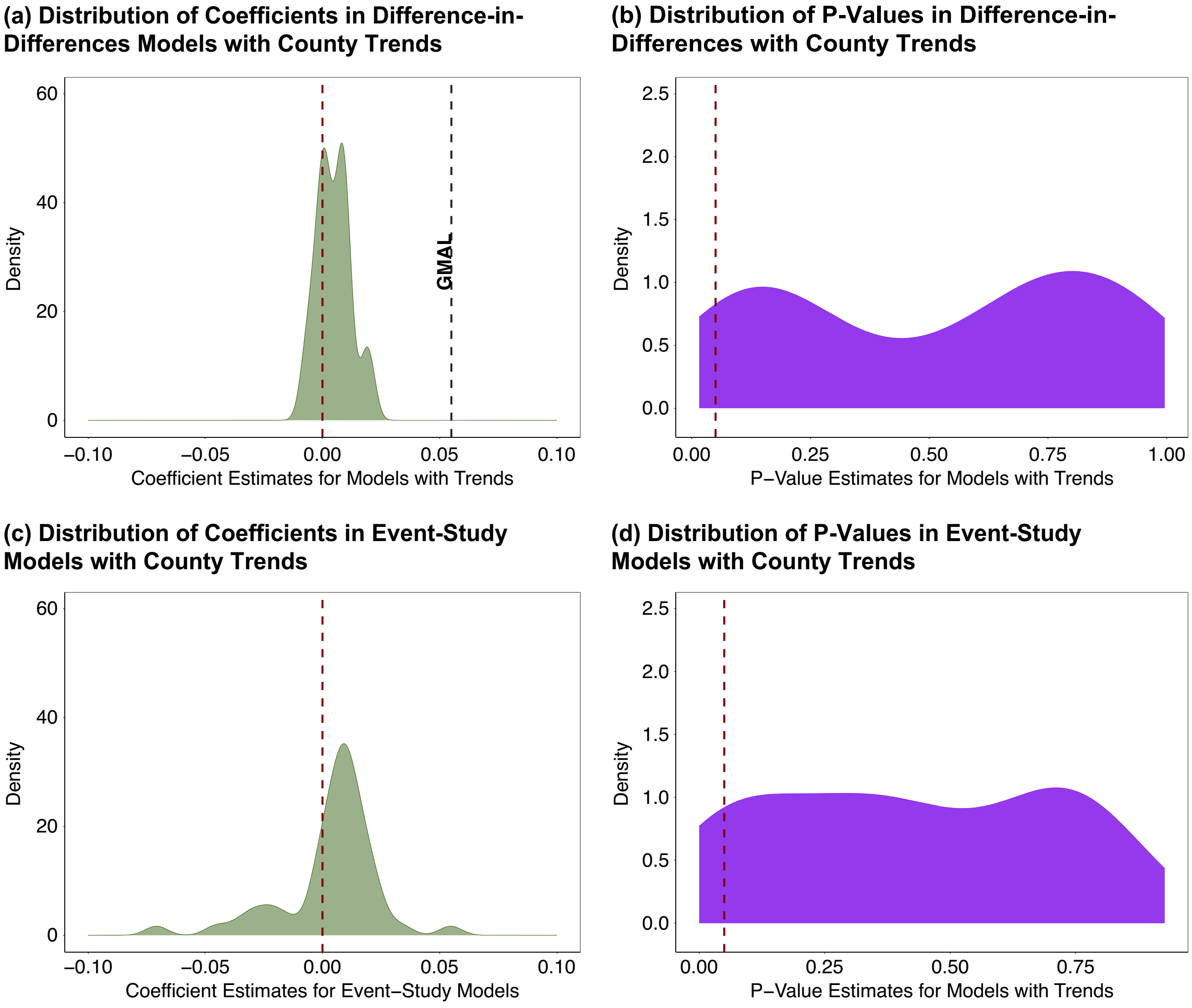
A few final words of guidance and caution remain. First, in all empirical checks, it is important not to forget theory. For example, we focus on whether shootings change Democratic vote shares in counties where they occur. We have done so because whether an effect occurs *there* is the central dispute. However, there are other potential effects. Perhaps shootings have spillover effects—with effects in adjacent counties—or as a function of the distance from a shooting, the time since a shooting, or the intensity of the shooting (e.g., the number of deaths/injuries). Alternatively, perhaps mass shootings have national effects. The best evidence currently suggests that none of these things occur (HHB).[46] However, scholars should remember there are often multiple ways of conceptualizing treatment exposure.

Second, it may, at times, be useful to unpack treatment effects at a more granular level—estimating, for example, the effect of individual shootings, rather than the average effect using the synthetic control method (Abadie, Diamond, and Hainmueller 2015;

[46] Only HHB consider these different types of coding treatment. They find null effects in all once accounting for time trends.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

FIGURE 11. Distribution of All Effect Estimates and *P*-Values for Models with County Trends



(a) Distribution of Coefficients in Difference-in-Differences Models with County Trends

(b) Distribution of P-Values in Difference-in-Differences with County Trends

(c) Distribution of Coefficients in Event-Study Models with County Trends

(d) Distribution of P-Values in Event-Study Models with County Trends

Note: Distribution of all model estimates with trends in Figure 5 in the first row and then for all the event-study estimates in the article on the bottom row. The event-study coefficients are shown for periods 0–4 post-treatment. The left panel in each row shows coefficients (in percentage point units). The right panel in each row shows the distribution of *p*-values across model specifications. *Takeaway:* Once we account for potential violations of parallel trends, the effects of shootings spike around zero, are only rarely significant, are not robust to slight changes in model specification, and are sometimes positive and sometimes negative.

Arkhangelsky et al. 2021; Kreif et al. 2016; Porreca 2022). Similarly, if interested in whether a subset of observations drives results, Broderick, Giordano, and Meager (2020) have an implementable procedure.[47]

Third, in some applications, implementing different modeling strategies for estimating treatment effects with longitudinal data may make sense. For example, others have attempted to bracket treatment effects of interest in longitudinal data. Specifically, Hasegawa, Webster, and Small (2019) recommend bracketing by splitting the control group into different groups based on relative comparisons to the treatment group. As they describe, "the basic idea is to consider one control group that has a lower expected outcome than the treated group in the before period and another control group that has a higher expected outcome than the treated group in the before period" (Hasegawa, Webster, and Small 2019, 372). They show estimators using the lower control group and the higher control group bracket the causal effect of treatment. This approach may be fruitful in cases where the pool of control units is large and easily split.[48] Likewise, Ding and Li (2019)

[47] See their zaminfluence*R* package.

[48] Implementing bracketing in our case, the bracket effects between 0.37 and 1.0 percentage points in the model with quadratic unit trends—still much smaller than prior estimates.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

and Angrist and Pischke (2008) suggest bracketing strategies combining the standard difference-in-differences specification with lagged dependent variables.[49,50]

Fourth, scholars may be interested in estimating distributional effects. Recent work combines difference-in-differences estimators with those quantile regression produces (see Callaway and Li 2019; Roth et al. 2022). Using this test, HHB show little signs of shifts at any point along the distribution of Democratic vote shares—that is, there is little evidence of polarizing effects in Democratic and Republican counties (see their Supplementary Figure A12 and surrounding discussion).

Finally, in making modeling decisions, one should acknowledge the tradeoffs between bias and precision and the importance of considering power in testing for pre-trends (Freyaldenhoven, Hansen, and Shapiro 2019; Roth 2022; Roth et al. 2022). For example, higher-order polynomials for unit-specific time trends require more power and may inflate standard errors. In our applied example, we have taken great care to pay attention to effect sizes, statistical significance, and the range of potential effects.

Overall, we note how to implement difference-in-differences designs depends, to a certain extent, on the nature of the data—that is, whether or not there are likely violations of parallel trends or unaccounted treatment effect heterogeneity or both. In the case of gun violence on electoral outcomes, the former appears to be key, while the latter less so. However, this may not always be the case. We think it is best to follow the suggestions outlined above to ensure inferences are not misleading. In our applied example, doing so reconciles why different studies using the same data arrived at vastly different conclusions regarding gun violence's effects on electoral vote shares.

SYNTHESIZING EVIDENCE FROM MULTIPLE SPECIFICATIONS

We note here a few important points required to come to a conclusion about mass shootings' effects on elections. In our particular example, there are many plausible models, a small number of which show statistically significant effects. Given the potential for bias and the role of researcher decisions, it is important for researchers to (1) address potential threats to inference outlined above, (2) be transparent about the effect of simple changes to model specification, and (3) take a "preponderance of evidence" rather than a "singular model" approach.[51]

What does this mean in the mass shootings context? Though on occasion we see intermittent statistically significant effects, these effects are (1) much smaller than previous research (i.e., GMAL and Yousaf) claims and (2) not robust to reasonable alternative specifications under the control of researchers. Taking a "singular model" approach makes researchers vulnerable to mistaken inferences given researcher degrees of freedom in choosing a model specification. However, a "preponderance of evidence" approach provides considerable reasons to doubt mass shootings have any significant, systematic, or large effects.[52] Models accounting for parallel-trends violations, while not completely ruling out some potential smaller effects (although almost all of these effects are not statistically significant), provide little to no evidence that mass shootings cause large electoral change in the United States and instead provide compelling evidence consistent with null effects.

This can be seen by synthesizing four pieces of evidence. First, though a very small number of corrected models show *much* smaller positive effect on Democratic vote share, almost all are not statistically significant. Second, negative effects also show up regularly across the small, but reasonable, changes to model specification within researcher control. Figure 11 plots the distribution of effects and *p*-values for all difference-in-differences models and event-study models we estimated above. As shown in panel a, all coefficients from difference-in-differences models with trends are much smaller than GMAL's TWFE. Some are positive and some are negative and the distribution spikes near zero. The average effect is 0.9 percentage points. Panel c shows that event-study models also spike at zero, with similar numbers of positive and negative effects. The average effect for all posttreatment periods is 0.4 percentage points and the average in the year immediately following treatment (i.e., period 0) is 0.07 percentage points. Third, when significant and positive effects do show up, these effects are not robust to slight variations in model specification within researchers' reasonable control. Fourth, sensitivity analyses embracing uncertainty around exact parallel-trends departures show results are *highly* sensitive to even minimal reasonable departures. Hence, the preponderance of evidence suggests large effects are implausible and that modest positive (or negative) effects are anything but sure.

[49] As Ding and Li (2019, 605) explain "for a true positive effect, if ignorability is correct, then mistakenly assuming parallel-trends will overestimate the effect; in contrast, if the parallel trends assumption is correct, then mistakenly assuming ignorability will underestimate the effect."

[50] For our case, the bracket is between 1 percentage point and 2.7 percentage points—much smaller than prior estimates. Moreover, this estimate is from the lagged dependent variable model without fixed effects. If we include the lagged dependent variable with fixed effects, the bounds are between 1 and 1.08 percentage points.

[51] In comparing the various approaches, leveraging new programming tools that make estimating multiple approaches at once easier may be useful, such as Hollenbach (2021).

[52] Arguing scholars should run many model specifications may prompt issues with multiple comparisons. Scholars should be careful not to overinterpret isolated significant effects in a deluge of otherwise insignificant results. However, properly adjusting for multiple comparisons across similar robustness checks is not well developed. Moreover, in arguing for the null, it is more conservative not to make any adjustments for multiple comparisons.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

CONCLUSION

In reconciling research on the effects of mass shootings on electoral outcomes, our work also highlights the considerations we argue should become standard practice given the hazards of navigating difference-in-differences designs. In addition to resolving an important question, we hope our article sparks a more nuanced approach to estimating difference-in-differences models. If appropriately used, the checks outlined above will help researchers make better inferences using this common identification strategy.

The methodological contribution we provide applies best to cases where the treatment may not be fully exogenous or may vary in timing across units, and with larger sample sizes having more cross sections than time points. Instances departing from these may leverage approaches similar to those outlined above, but with unique features. Moreover, we have not explored some valuable aspects of panel data estimation recently developed for scenarios with very few treatment units (e.g., synthetic controls) that are valuable. Finally, our work is applied to a context where there is not currently, nor any prospect of a future, experimental baseline. While the econometric literature has long highlighted the value of the checks we run—through proofs, simulations, and other validation techniques—there is not yet (to the best of our knowledge) a comparison of difference-in-differences tools in our arsenal to a randomized baseline. Future work would do well to find other contexts where randomization is possible and add this benchmarking task to our suite of studies on this widely used method, as has been done with other methodological techniques (e.g., Arceneaux, Gerber, and Green 2006; Green et al. 2009).

Returning to the context of this study, America's legacy of gun violence is heartbreaking and the thousands of deaths that occur from guns each year constitute a policy failure of epic proportions. Yet scholars have disagreed whether policymakers relative inaction occurs in spite of (or as a result of a lack of) an electoral response. While agreeing mass shootings do not effect voter turnout, scholars have come to vastly different conclusions about mass shootings' effects on vote shares. We show that we cannot definitively conclude gun violence has any impact on Democratic vote shares (either positive or negative) and that previous work showing such relationships fails to navigate many of the pitfalls of difference-in-differences designs, specifically a failure of the parallel trends assumption. Moreover, even the most generous interpretation—i.e., entirely ignoring statistical uncertainty (something that prior work suggests should not be done; see, e.g., Stock and Watson 2020)—suggests shootings have, at best, modest effects on Democratic vote share *much* smaller than prior research emphasizes. Looking across all robustness checks, we cannot conclude mass shootings of any kinds substantially affect election outcomes. Such a conclusion comes only from selecting highly sensitive results that are not robust.

Furthermore, these estimates are all *local* to the county in which the shooting occurred. Though all mass shootings are repugnant, they are (thankfully) relatively rare. Given shootings only occurred in 0.4% of counties (116 total; 11.6 per election) in HHB's data, 0.4% of counties (115 total; 11.5 per election) in GMAL's data, and only 0.5% of counties (72 total; 14.4 per election) in Yousaf's data further emphasizes the limited impact shootings have on elections.[53] Ultimately, both the modest effect sizes in percentage points and their limited scope show mass shootings have little substantive consequence for election outcomes. *Even if* we take point estimates above at face value and ignore statistical uncertainty (something we certainly should *not* do), given the county-specific effect of the size observed and their infrequency mass shootings would have *virtually no effect on any statewide or national election.*

Lastly, as a reminder, there are often multiple ways of conceptualizing treatment exposure. These include, but are not limited to, short-lived treatments (constrained only to the period when they happen) or longer-term treatments (turned on in all periods after treatment occurs), spillover treatments (affecting units adjacent to treatment), varying dosage treatments, and even the possibility of national treatments drowning out any potential local effects. However, the best evidence we have currently suggests that *none* of these occur in mass shootings contexts (see HHB, 1377).

Our work sets the table for future work on the political economy of gun violence and retrospective voting/accountability more generally. Future work would do well to explore *why* mass shootings fail to substantively change the electoral incentives elected officials face, despite having favorable conditions to do so (HHB). Our work also acts a guide for researchers navigating the potential pitfalls of difference-in-difference designs. Future work would do well to continue to advance the boundaries of this promising and increasingly common method for making causal inferences.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055424000108.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/GH69TI.

ACKNOWLEDGMENTS

We would like to thank Laura García-Montoya, Ana Arjona, Matthew Lacombe, and Hasin Yousaf for making their replication data and code available. We also

[53] Only HHB consider whether there are spillover effects on adjacent counties, and find none (see 1377).

thank Matt Baldwin, Charles Crabtree, Bernard Fraga, Andrew Goodman-Bacon, Justin Grimmer, Zoltan Hajnal, Nazita Lajevardi, Gareth Nellis (and the UCSD Graduate Student Seminar), Ben Newman, Tyler Reny, Brian Shaffner, Zayne Sember, Neil Visalvanich, Bertrand Wilden, and Yiqing Xu, and five anonymous reviewers for their feedback.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human participants.

REFERENCES

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. “When Should You Adjust Standard Errors for Clustering?” *Quarterly Journal of Economics* 138 (1): 1–35.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. “Comparative Politics and the Synthetic Control Method.” *American Journal of Political Science* 59 (2): 495–510.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics.” *Journal of Economic Perspectives* 24 (2): 3–30.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. “Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment.” *Political Analysis* 14 (1): 37–62.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. “Synthetic Difference-in-Differences.” *American Economic Review* 111 (12): 4088–118.

Armitage, Seth. 1995. “Event Study Methods and Evidence on Their Performance.” *Journal of Economic Surveys* 9 (1): 25–52.

Bai, Jushan, and Serena Ng. 2021. “Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data.” *Journal of the American Statistical Association* 116 (536): 1746–63.

Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. “How Much Should We Trust Staggered Difference-in-Differences Estimates?.” *Journal of Financial Economics* 144 (2): 370–95.

Barney, David J., and Brian F. Schaffner. 2019. “Reexamining the Effect of Mass Shootings on Public Support for Gun Control.” *British Journal of Political Science* 49 (4): 1555–65.

Beck, Nathaniel L., Jonathan N. Katz, and Umberto G. Mignozzetti. 2014. “Of Nickell Bias and Its Cures: Comment on Gaibulloev, Sandler, and Sul.” *Political Analysis* 22 (2): 274–8.

Berge, Laurent, Sebastian Krantz, and Grant McDermott. 2022. “Fixest: Fast Fixed-Effects Estimations.” *CRAN*.

Bilinski, Alyssa, and Laura A. Hatfield. 2018. “Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions.” Preprint, arXiv:1805.03273.

Binder, John. 1998. “The Event Study Methodology since 1969.” *Review of Quantitative Finance and Accounting* 11 (2): 111–37.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. “Revisiting Event Study Designs: Robust and Efficient Estimation.” Preprint, arXiv:2108.12419.

Broderick, Tamara, Ryan Giordano, and Rachael Meager. 2020. “An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?” Preprint, arXiv:2011.14999.

Callaway, Brantly, and Tong Li. 2019. “Quantile Treatment Effects in Difference in Differences Models with Panel Data.” *Quantitative Economics* 10 (4): 1579–618.

Callaway, Brantly, and Pedro H. C. Sant’Anna. 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225 (2): 200–30.

Cameron, A. Colin, and Douglas L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50 (2): 317–72.

Clarke, Damian, and Kathya Tapia-Schythe. 2021. “Implementing the Panel Event Study.” *Stata Journal* 21 (4): 853–84.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press.

De Chaisemartin, Clément, and Xavier d’Haultfoeuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110 (9): 2964–96.

De Chaisemartin, Clément, and Xavier D’Haultfoeuille. 2023. “Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey.” *The Econometrics Journal* 26 (3): C1–C20.

de Chaisemartin, Clément, Xavier D’Haultfoeuille, and Yannick Guyonvarch. 2019. “DID_MULTIPLEGT: Stata Module to Estimate Sharp Difference-in-Difference Designs with Multiple Groups and Periods.” Statistical Software Components S458643, Boston College Department of Economics.

DeSilver, Drew. 2016. “The Growing Democratic Domination of Nation’s Largest Counties.” *Pew Research Center*, July 21.

Ding, Peng, and Fan Li. 2019. “A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment.” *Political Analysis* 27 (4): 605–15.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse Shapiro. 2022. “Xtevent: Stata Module to Estimate and Visualize Linear Panel Event-Study Models.” Working Paper.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. 2021. “Visualization, Identification, and Estimation in the Linear Panel Event-Study Design.” Working Paper, National Bureau of Economic Research.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. “Pre-Event Trends in the Panel Event-Study Design.” *American Economic Review* 109 (9): 3307–38.

García-Montoya, Laura, Ana Arjona, and Matthew Lacombe. 2022. “Violence and Voting in the United States: How School Shootings Affect Elections.” *American Political Science Review* 116 (3): 807–26.

Goodman-Bacon, Andrew. 2019. “So You’ve Been Told to Do My Difference-in-Differences Thing: A Guide.” Working Paper, Vanderbilt University.

Goodman-Bacon, Andrew. 2021. “Difference-in-Differences with Variation in Treatment Timing.” *Journal of Econometrics* 225 (2): 254–77.

Goodman-Bacon, Andrew, Thomas Goldring, and Austin Nichols. 2019. “BACONDECOMP: Stata Module to Perform a Bacon Decomposition of Difference-in-Differences Estimation.” Statistical Software Components S458676, Boston College Department of Economics.

Goss, Kristin. 2010. *Disarmed: The Missing Movement for Gun Control in America*. Princeton, NJ: Princeton University Press.

Green, Donald P., Terence Y. Leong, Holger L. Kern, Alan S. Gerber, and Christopher W. Larimer. 2009. “Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks.” *Political Analysis* 17 (4): 400–17.

Grimmer, Justin, Eitan Hersh, Marc Meredith, Jonathan Mummolo, and Clayton Nall. 2018. “Obstacles to Estimating Voter ID Laws’ Effect on Turnout.” *Journal of Politics* 80 (3): 1045–51.

Hansen, Ben B., and Jake Bowers. 2008. “Covariate Balance in Simple, Stratified and Clustered Comparative Studies.” *Statistical Science* 23(2): 219–36.

Hartman, Erin, and F. Daniel Hidalgo. 2018. “An Equivalence Approach to Balance and Placebo Tests.” *American Journal of Political Science* 62 (4): 1000–13.

Hartman, Todd K., and Benjamin J. Newman. 2019. “Accounting for Pre-Treatment Exposure in Panel Data: Re-Estimating the Effect of Mass Public Shootings.” *British Journal of Political Science* 49 (4): 1567–76.

Hasegawa, Raiden B., Daniel W. Webster, and Dylan S. Small. 2019. “Evaluating Missouri’s Handgun Purchaser Law: A Bracketing

Method for Addressing Concerns about History Interacting with Group." *Epidemiology* 30 (3): 371–9.

Hassell, Hans J. G., John B. Holbein, and Matthew Baldwin. 2020. "Mobilize for Our Lives? School Shootings and Democratic Accountability in US Elections." *American Political Science Review* 114 (4): 1375–85.

Healy, Andrew, and Gabriel S. Lenz. 2017. "Presidential Voting and the Local Economy: Evidence from Two Population-Based Data Sets." *Journal of Politics* 79 (4): 1419–32.

Holbein, John B., and Hans J. G. Hassell. 2024. "Replication Data for: Navigating Potential Pitfalls in Difference-in-Differences Designs: Reconciling Conflicting Findings on Mass Shootings' Effect on Electoral Outcomes." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/GH69TI.

Hollenbach, Florian. 2021. "Comparing Staggered DiD." https://github.com/fhollenbach/did_compare/blob/main/ComparingDiD.md.

Kahn-Lang, Ariella, and Kevin Lang. 2020. "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications." *Journal of Business & Economic Statistics* 38 (3): 613–20.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23 (3): 313–35.

Keele, Luke, and William Minozzi. 2013. "How Much Is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data." *Political Analysis* 21 (2): 193–216.

Kreif, Noémi, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. 2016. "Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units." *Health Economics* 25 (12): 1514–28.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2024. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science* 68 (1): 160–76.

Luca, Michael, Deepak Malhotra, and Christopher Poliquin. 2020. "The Impact of Mass Shootings on Gun Policy." *Journal of Public Economics* 181: 104083.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. 2023. "Testing for the Appropriate Level of Clustering in Linear Regression Models." *Journal of Econometrics* 235 (2): 2027–56.

Manski, Charles F., and John V. Pepper. 2018. "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions." *Review of Economics and Statistics* 100 (2): 232–44.

Marcus, Michelle, and Pedro H. C. Sant'Anna. 2021. "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics." *Journal of the Association of Environmental and Resource Economists* 8 (2): 235–75.

Marsh, Wayde Z. C. 2022. "Trauma and Turnout: The Political Consequences of Traumatic Events." *American Political Science Review* 117 (3): 1036–52.

Mou, Hongyu, Licheng Liu, and Yiqing Xu. 2022a. "Package 'panelView.'"

Mou, Hongyu, Licheng Liu, and Yiqing Xu. 2022b. "panelView: Panel Data Visualization in R and Stata." SSRN 4202154.

Musu-Gillette, Lauren, Anlan Zhang, Ke Wang, Jana Kemp, Melissa Diliberti, and Barbara A. Oudekerk. 2018. "Indicators of School Crime and Safety: 2017." Report, National Center for Educational Statistics (NCES 2018-036).

Montiel Olea, José Luis, and Mikkel Plagborg-Møller. 2019. "Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs." *Journal of Applied Econometrics* 34 (1): 1–17.

Porreca, Zachary. 2022. "Synthetic Difference-in-Differences Estimation with Staggered Treatment Timing." *Economics Letters* 220: 110874.

Rambachan, Ashesh, and Jonathan Roth. 2021. "An Honest Approach to Parallel Trends." Unpublished Manuscript, Harvard University.

Rambachan, Ashesh, Jonathan Roth, and Mauricio Caceres Bravo. 2021. "HonestDiD." https://github.com/asheshrambachan/HonestDiD/tree/bc576d5f338dd01ebb641f30215792c0605bc08a.

Rogowski, Jon C., and Patrick D. Tucker. 2019. "Critical Events and Attitude Change: Support for Gun Control after Mass Shootings." *Political Science Research and Methods* 7 (4): 903–11.

Rossin-Slater, Maya, Molly Schnell, Hannes Schwandt, Sam Trejo, and Lindsey Uniat. 2020. "Local Exposure to School Shootings and Youth Antidepressant Use." *Proceedings of the National Academy of Sciences* 117 (38): 23484–9.

Roth, Jonathan. 2022. "Pre-Test with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4(3): 305–22.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics* 235 (2): 2218–44.

Schmidheiny, Kurt, and Sebastian Siegloch. 2023. "On Event Studies and Distributed Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization." *Journal of Applied Econometrics* 38(5), 695–713.

Sides, John, Lynn Vavreck, and Christopher Warshaw. 2022. "The Effect of Television Advertising in United States Elections." *American Political Science Review* 116 (2): 702–18.

Stock, James H., and Mark W. Watson. 2020. *Introduction to Econometrics*. New York: Pearson.

Sun, Liyang. 2021. "EVENTSTUDYINTERACT: Stata Module to Implement the Interaction Weighted Estimator for an Event Study." Statistical Software Components S458978, Boston College Department of Economics.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.

U.S. Government Accountability Office. 2020. "K-12 Education: Characteristics of School Shootings." GAO-20-455.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39: 453–69.

Xu, Yiqing. 2023. "Causal Inference with Time-Series Cross-Sectional Data: A Reflection." In *Oxford Handbook of Engaged Methodological Pluralism in Political Science*, Vol. 1, eds. Janet M. Box-Steffensmeier, Dino P. Christenson, and Valeria Sinclair-Chapman, online edition. https://doi.org/10.1093/oxfordhb/9780192868282.013.30.

Yousaf, Hasin. 2021. "Sticking to One's Guns: Mass Shootings and the Political Economy of Gun Control in the United States." *Journal of the European Economic Association* 19 (5): 2765–802.

Zhang, Shuo. 2022. "DIDmultiplegt: Estimation in DID with Multiple Groups and Periods." CRAN.

https://doi.org/10.1017/S0003055424000108 Published online by Cambridge University Press

Navigating Potential Pitfalls in Difference-in-Differences Designs: Reconciling Conflicting Findings on Mass Shootings' Effect on Electoral Outcomes - Online Appendix

List of Tables

List of Figures

1 Commonalities in the Papers in Question

We think it important to note that all three papers in qustion have some common findings. GMAL, Yousaf, and HHB all find no effects of mass shootings on voter turnout (as we show below, turnout does not appear to have the trend differences that plague Democratic vote share—see Figure S10.) This overall finding is also corroborated by a recent paper by Marsh (2022), who finds that changes in turnout after mass shootings are "not statistically distinguishable from zero." While Marsh does provide some evidence that mass shootings close to an election have a slight positive effect on turnout (see Figure 1, Marsh 2022), HHB show that the effects of school shootings close to an election on turnout are highly sensitive to model specification (see HHB, Figure A7) a pattern also somewhat evident in Marsh's models (see Marsh (2022), Table E2 and E5)). Importantly, then, given the lack of any substantive effect on turnout, any increase in Democratic vote share should come from persuasion, rather than mobilization, unless gun violence simultaneously demobilizes Republicans and mobilizes Democrats *at the exact same rates*, which is highly unlikely. However, any persuasive effect would also likely show up in attitudinal shifts and previous research on the attitudinal effects of mass shootings has disagreed whether attitudinal effects are present and, if they are, whether these effects are polarizing or a uniform leftward shift (Barney and Schaffner 2019; Hartman and Newman 2019; Rogowski and Tucker 2019). An absence of an attitudinal shift does not alone undermine GMAL and Yousaf's results, but it provides a theoretical reason to question their results. Ultimately, however, our goal here is to try and settle the first-order question of whether gun violence has any effect on vote shares in the communities in which they happen. If there was, we could then proceed to adjudicate between mobilization and persuasion mechanisms. As we show, however, there is virtually no clear support for an effect on vote shares.

Table S1: Differences in All Studies on the Effects of Gun Violence on Electoral Vote Shares

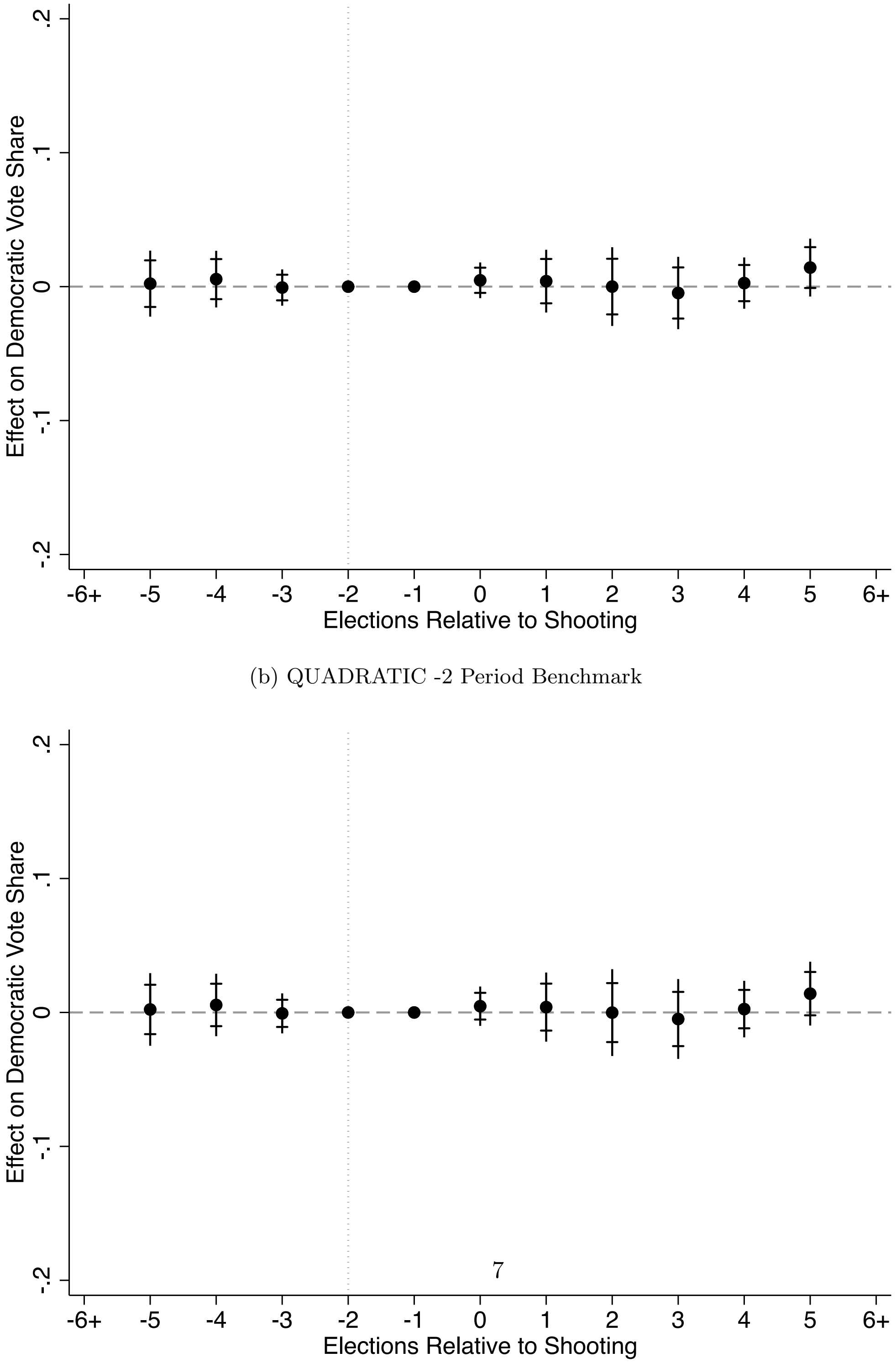
		GMAL	**Yousaf**	**HHB**
Data	Shootings	"Rampage-style" school shootings: "Rampage-style" shootings are shootings that "take place on a school-related public stage before an audience; involve multiple victims, some of whom are shot simply for their symbolic significance or at random; involve one or more shooters who are students or former students of the school and where the motivation of the shooting [does not] correlate with gang violence or targeted militant or terroristic activity" (GLAM, 1)	Mass Shootings: Mass shootings are all shootings "leading to four or more deaths at one location" (Yousaf, 2770)	All school shootings (HHB, 1377)
	Years	1980 to 2016	2000 to 2016	2000 to 2018
	Vote Outcomes	Presidential election returns only	Presidential, gubernatorial, senatorial, and congressional election returns from presidential election years only	Presidential, congressional, state, and local election returns in all years
Methods	Model Specifications	Difference-in-Differences TWFE **(No county specific time-trends included)**	Difference-in-Differences TWFE **(No county specific time-trends included)**	Difference-in-Differences TWFE with county specific time-trends
	Standard Errors	Clustered at the state level	Clustered at the state level	Clustered at the county (treatment) level

2 Solutions for Effect Heterogeneity Problems

Three other solutions to treatment effect heterogeneity problems identified in the literature are worth mentioning. The differences between these are nuanced and not all may be well-suited in some applications. First, like Sun and Abraham (2020), Callaway and Sant'Anna (2021) argue that scholars should use a method that restricts to "clean comparisons" applying to scenarios where "(i) multiple time periods, (ii) variation in treatment timing, and (iii) when the 'parallel-trends assumption' holds potentially only after conditioning on observed covariates." This approach facilitates the estimation of propensity scores conditional on observed covariates to help achieve pre-treatment balance. With this approach, we make the panel balanced and code all post-treatment units as treated as doing so is more appropriate for this approach. We show this approach in Figure S19. Unfortunately, this approach has limited value in our application for two reasons. First, *even when* one uses "clean comparions" as suggested by Callaway and Sant'Anna (2021) and covariates, differential pre-treatment trends issues remain.[1] Second, their approach does not yet extend to models with unit-specific time trends. These may be less of an issue in other applications, so we include these as an illustration of this method and its results. Second, De Chaisemartin and d'Haultfoeuille (2020) provide an alternate approach for assessing and addressing implemented in the `did_multiplegt` package in *STATA* and `DIDmultiplegt` package in *R*. With this approach, we code all post-treatment units as treated as doing so is more appropriate for this approach. Unfortunately, this approach only allows linear trends and doesn't allow for flexibility in other model parameters that approaches like other methods afford and is more computationally intensive. Still, to illustrate this method, we provide the results for this in Tables S8 and S9. Finally, Borusyak et al. (2021, 1) use "an intuitive 'imputation' form [where] treatment-effect heterogeneity is unrestricted." This approach is implemented in the `did_imputation` package in *STATA* and `didimputation` package in *R*. We use the treatment of coding treatment only in the current period as it is more appropriate to do so for this approach. When we implement this approach, we still see a sizable effect both pre- and post-treatment in the TWFE. Unfortunately, this approach is not currently designed to implement with unit specific trends in our example. Though the package technically does allow trends, the help file warns users to "Use [trends] with caution: the command may not recognize that imputation is not possible for some treated observations." This appears to be the case in our application.

[1] Relevant packages here include the `csdid` in in *STATA* and in *R* and *hdidregress* and *xthdidregress* in *STATA* 18.

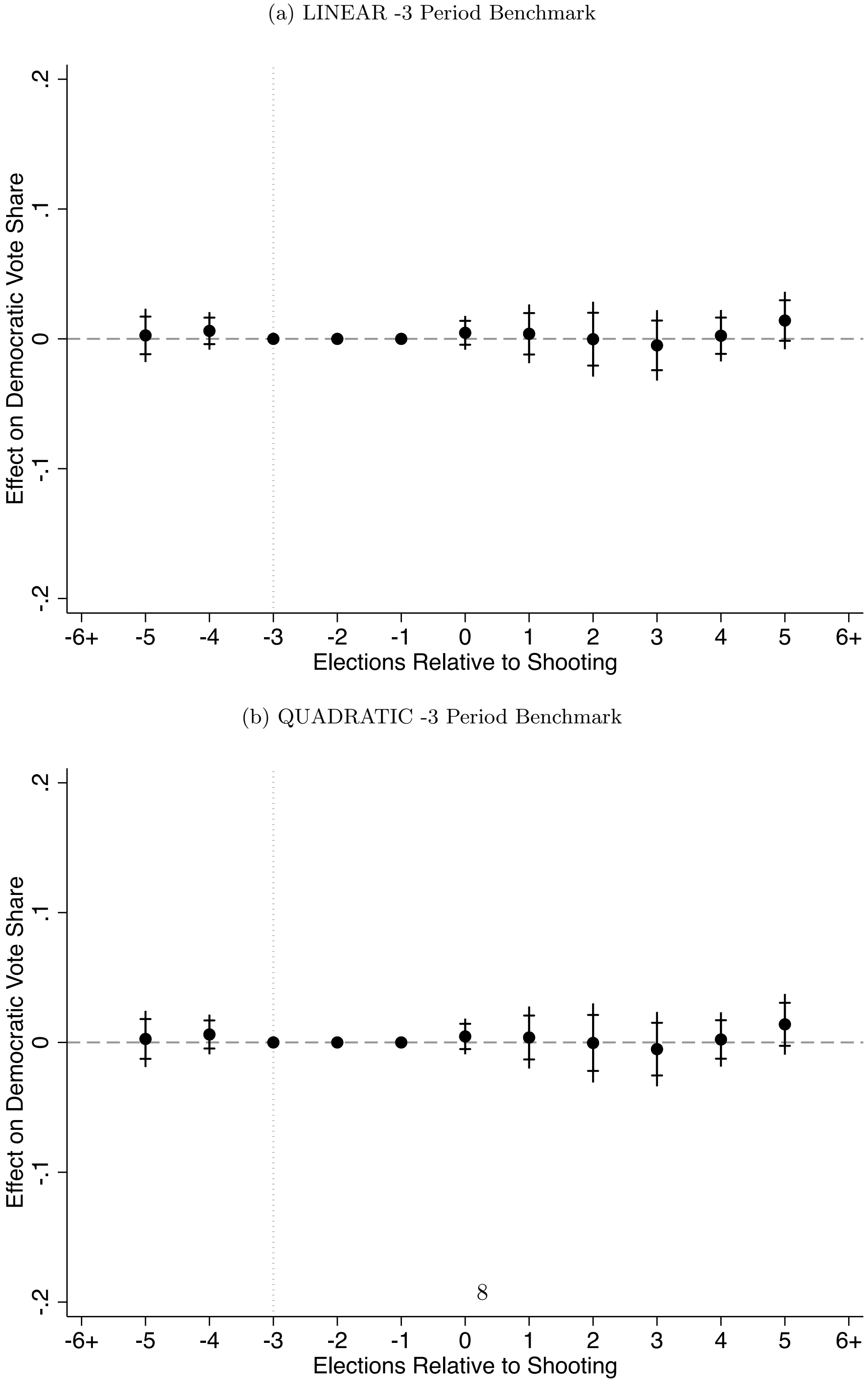
Figure S1: Results Alternate Baseline Periods in Event Study Design that Accounts for County Specific Time Trends


(a) LINEAR -2 Period Benchmark

(b) QUADRATIC -2 Period Benchmark

Figure shows the results from using other pre-treatment periods as the baseline as suggested by Freyaldenhoven et al. (2021). **Takeaway:** Benchmarked to pre-treatment trends at t-2, the estimates are even smaller, and even less suggestive of mass shootings having an effect on electoral outcomes.

Figure S2: Results Alternate Baseline Periods in Event Study Design that Accounts for County Specific Time Trends (cont'd)



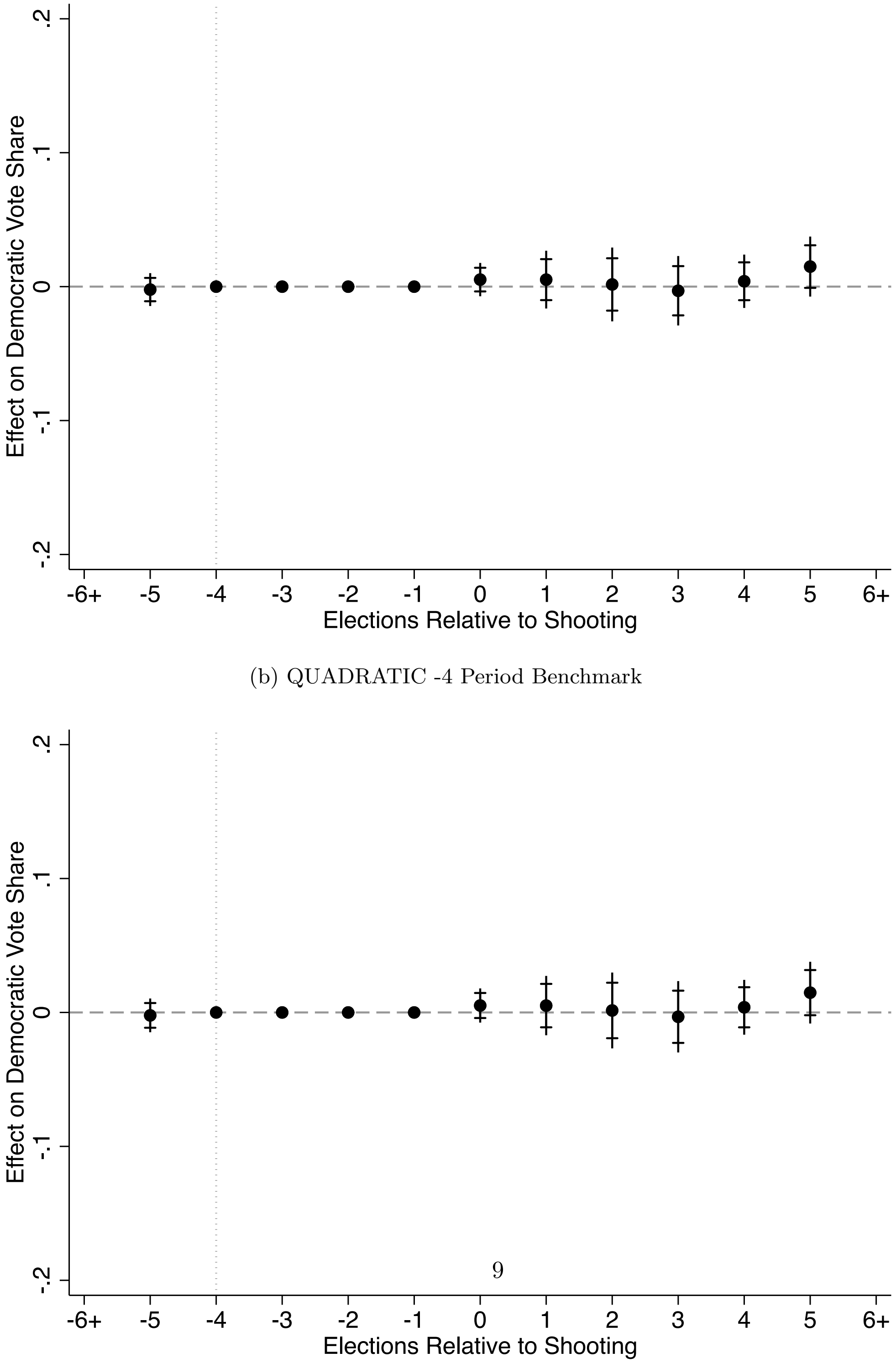
(a) LINEAR -3 Period Benchmark

(b) QUADRATIC -3 Period Benchmark

Figure shows the results from using other pre-treatment periods as the baseline as suggested by Freyaldenhoven et al. (2021). **Takeaway:** Benchmarked to pre-treatment trends at t-3, the estimates are even smaller, and even less suggestive of mass shootings having an effect on electoral outcomes.

Figure S3: Results Alternate Baseline Periods in Event Study Design that Accounts for County Specific Time Trends (cont'd)

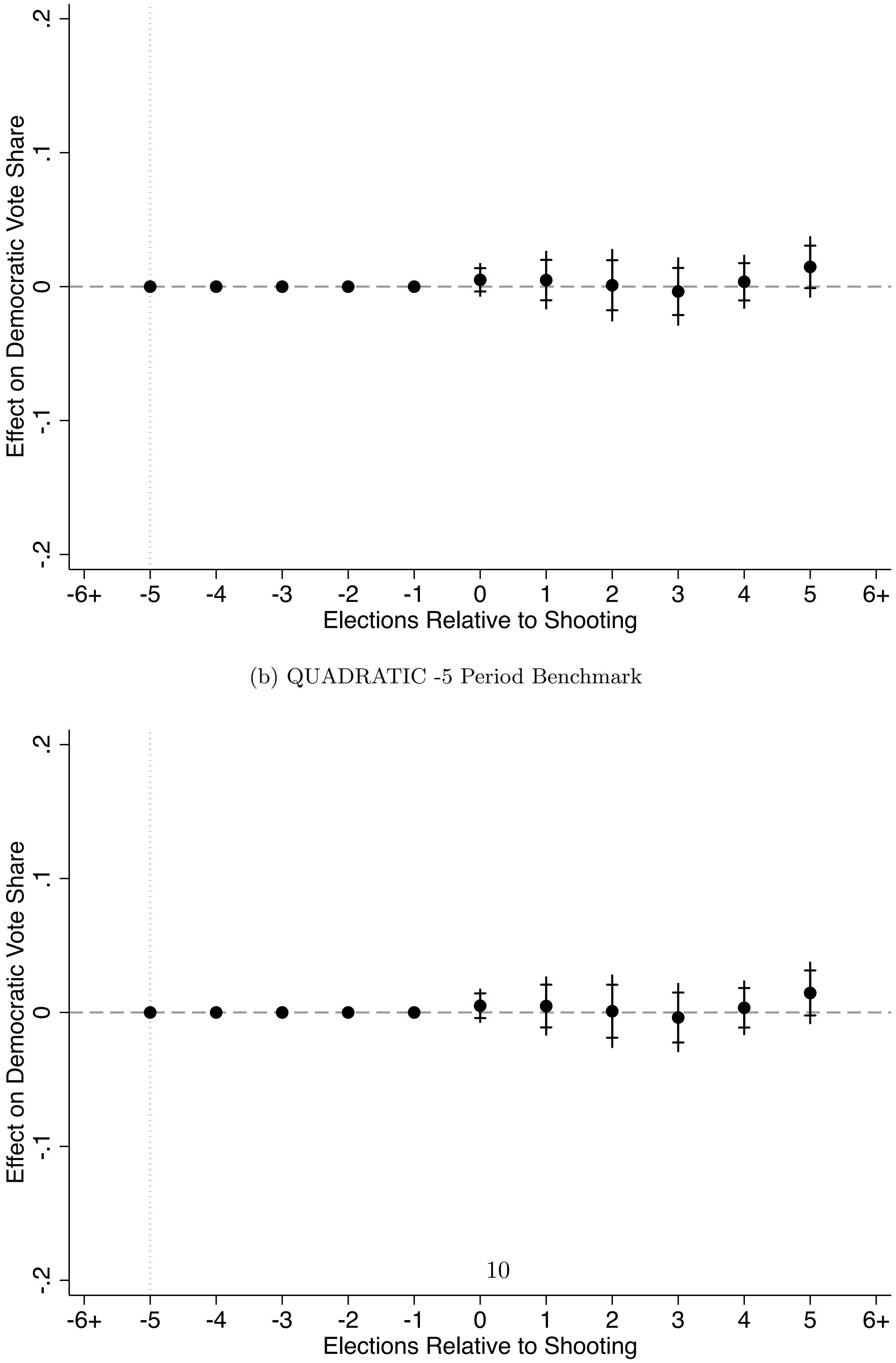
(a) LINEAR -4 Period Benchmark



(b) QUADRATIC -4 Period Benchmark

Figure shows the results from using other pre-treatment periods as the baseline as suggested by Freyaldenhoven et al. (2021).**Takeaway:** Benchmarked to pre-treatment trends at t-4, the estimates are even smaller, and even less suggestive of mass shootings having an effect on electoral outcomes.

Figure S4: Results Alternate Baseline Periods in Event Study Design that Accounts for County Specific Time Trends (cont'd)

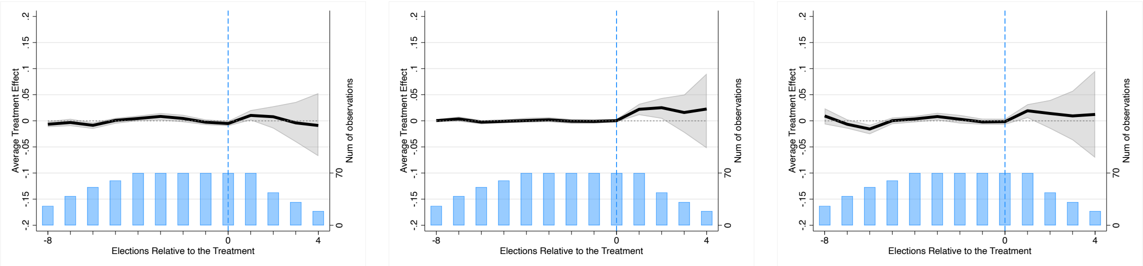


(a) LINEAR -5 Period Benchmark

(b) QUADRATIC -5 Period Benchmark

Figure shows the results from using other pre-treatment periods as the baseline as suggested by Freyaldenhoven et al. (2021).**Takeaway:** Benchmarked to pre-treatment trends at t-5, the estimates are even smaller, and even less suggestive of mass shootings having an effect on electoral outcomes.

Figure S5: Interactive Fixed Effects Counterfactual Estimator



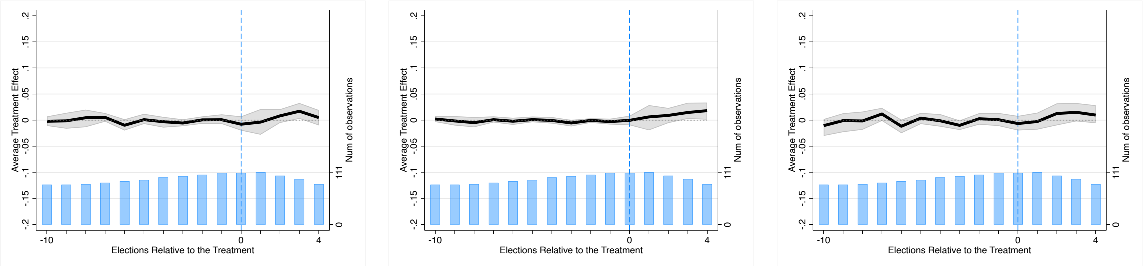
(a) Interactive Fixed Effects, r=2, degree=3

(b) Interactive Fixed Effects, r=3, degree=3

(c) Interactive Fixed Effects, r=1, degree=3

Figure shows the results from using Interactive Fixed Effects Counterfactual Estimator developed by Liu et al. (2021) with different values of r—the number of factors used in estimation—and the integer specifying the order of the polynomial trend term. **Takeaway:** In the interactive fixed effects models, there is no evidence of the substantial effects shown in more simplistic model specifications that do not account for potential violations of the parallel-trends assumption.

Figure S6: Interactive Fixed Effects Counterfactual Estimator (2)



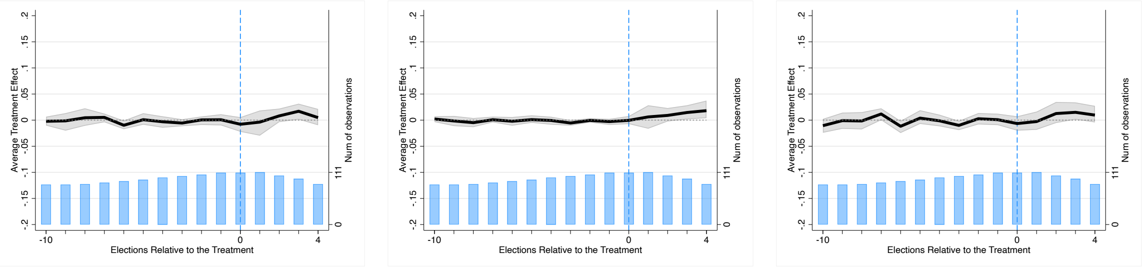
(a) Interactive Fixed Effects, r=2, degree=2

(b) Interactive Fixed Effects, r=3, degree=2

(c) Interactive Fixed Effects, r=1, degree=2

Figure shows the results from using Interactive Fixed Effects Counterfactual Estimator developed by Liu et al. (2021) with different values of r—the number of factors used in estimation—and the integer specifying the order of the polynomial trend term. **Takeaway:** In the interactive fixed effects models, there is no evidence of the substantial effects shown in more simplistic model specifications that do not account for potential violations of the parallel-trends assumption.

Figure S7: Interactive Fixed Effects Counterfactual Estimator (3)



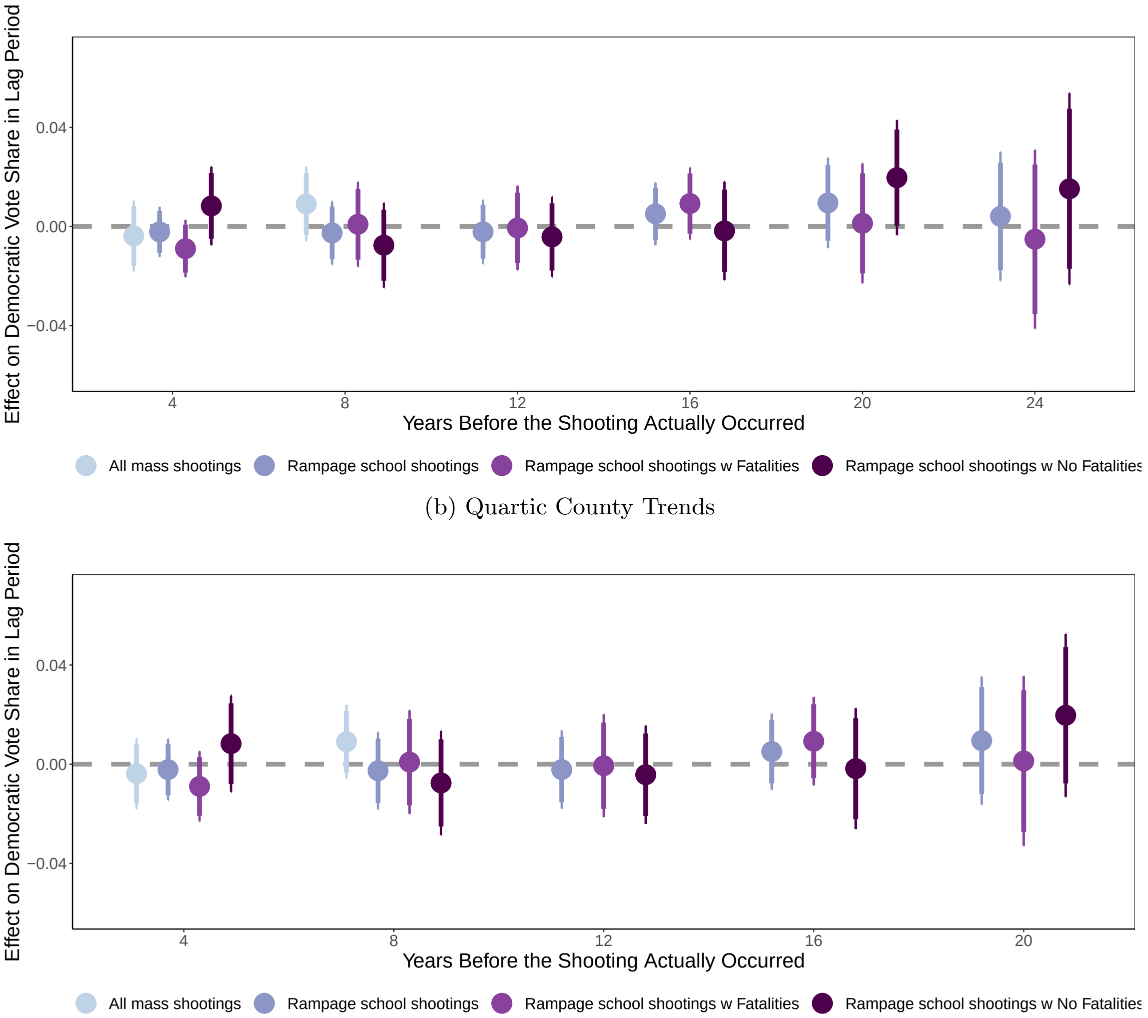
(a) Interactive Fixed Effects, r=2, degree=4

(b) Interactive Fixed Effects, r=3, degree=4

(c) Interactive Fixed Effects, r=1, degree=4

Figure shows the results from using Interactive Fixed Effects Counterfactual Estimator developed by Liu et al. (2021) with different values of r—the number of factors used in estimation—and the integer specifying the order of the polynomial trend term. **Takeaway:** In the interactive fixed effects models, there is no evidence of the substantial effects shown in more simplistic model specifications that do not account for potential violations of the parallel-trends assumption.

Figure S8: Pre-Treatment Effects with Alternate County-Specific Trend Types

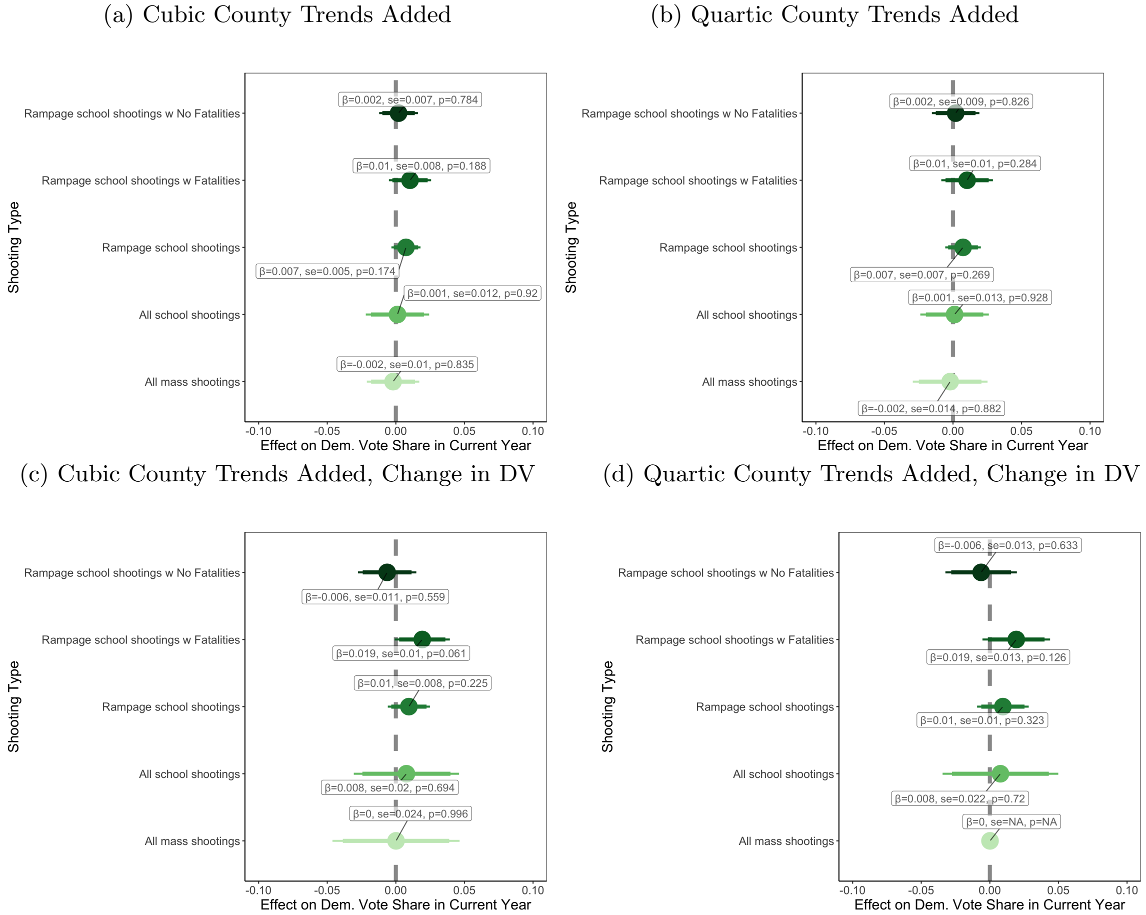


(a) Cubic County Trends

(b) Quartic County Trends

Figure shows the results from using higher order polynomial functional forms for the county-specific trends. The cubic trends model omits the 28 year lag because there are not enough observations in the GMAL data to estimate a model with this many high dimensional fixed effects. The quartic trends model omits the 24 and 28 year lag for the same reason. **Takeaway:** In contrast to the TWFE estimates shown in panels (a) and (b) in Figure 3 in the text (but consistent with specifications with linear and quadratic time trends), specifications with cubic and quadratic time trends show balance prior to when the shooting occurred.

Figure S9: The Effect of Mass Shootings on Presidential Election Returns Once County-Specific Trends are Absorbed, Alternate Polynomial Orders



(a) Cubic County Trends Added

(b) Quartic County Trends Added

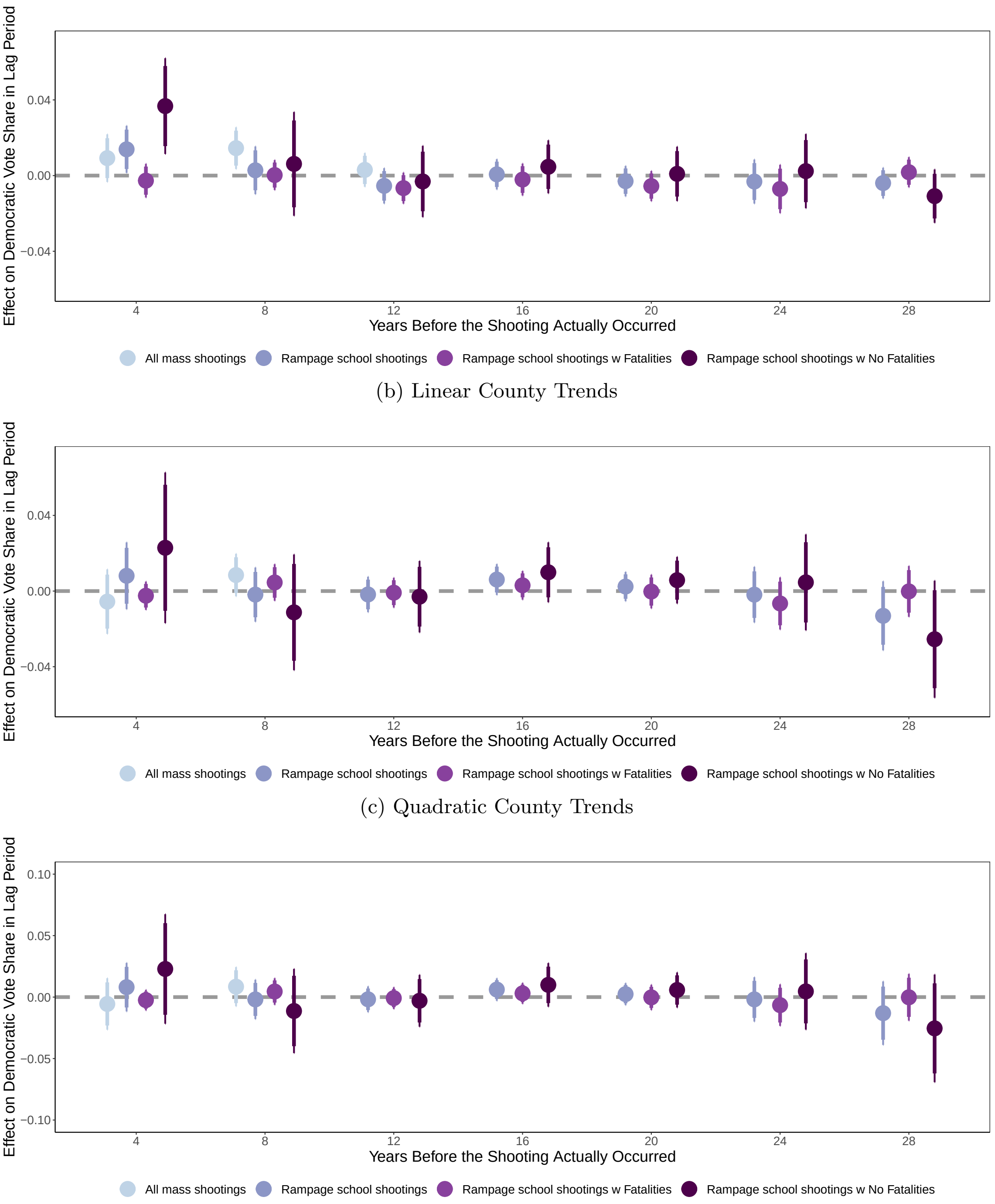
(c) Cubic County Trends Added, Change in DV

(d) Quartic County Trends Added, Change in DV

Figure shows the effect of mass shootings of various types once we account for differential trends in Democratic vote share across counties in the United States—this time with cubic and quartic county-specific trends. Within each panel, the first 3 estimates are using the GMAL coding of mass shootings and their data, the next comes from HHB, and the last comes from Yousaf. The upper left panel shows specifications with cubic county trends, the upper right panel shows specifications with quartic county trends, the bottom left panel shows specifications with cubic county trends and using a change in Democratic vote share over the prior 4-year-previous election, the bottom right panel shows specifications with quartic county trends and using a change in Democratic vote share over the prior 4-year-previous election. In the last panel, the standard error will not estimate for the Yousafdata as there are not observations in this shorter time series to do so. Coefficients, standard errors, and p-values are labeled for each coefficient. **Takeaway:** Once we account for differential trends across counties, the effects of mass shootings—be they located on school grounds or not, or be they rampage style or not—are all small and precisely-estimated.

Figure S10: Pre-Treatment Effects on Turnout

(a) TWFE

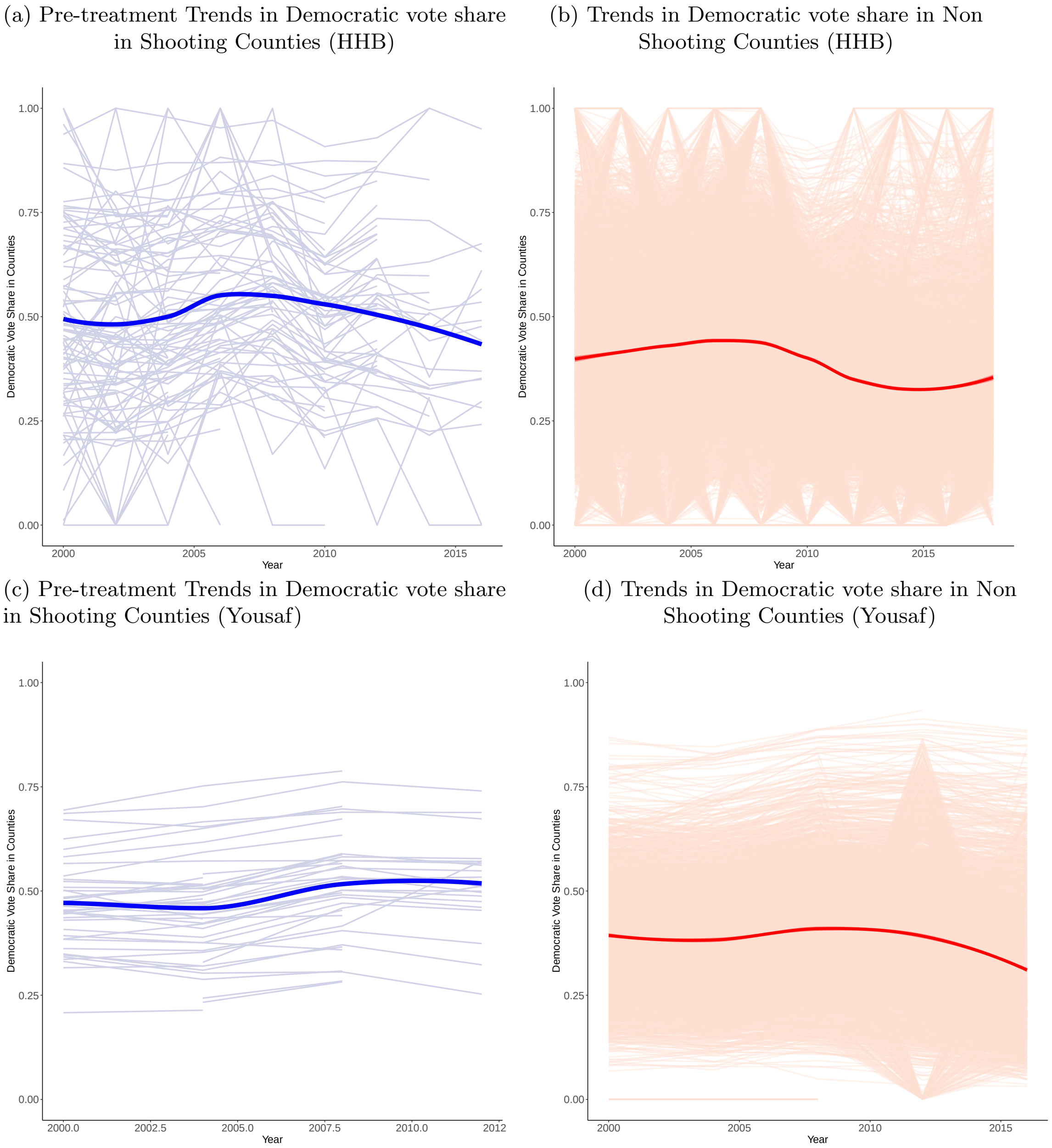


(b) Linear County Trends

(c) Quadratic County Trends

Figure shows the effect of mass shootings on voter turnout in the years prior to when a shooting occurred. **Takeaway:** In contrast to the effects of mass shootings on Democratic vote share which is plagued by trend differences pre-treatment, turnout does not appear to suffer from the same problem, as there is balance pre-treatment.

Figure S11: Trends in Presidential Vote Share in Counties With Mass Shootings Prior to These Shootings Occurring, Compared to Trends in Counties Without a Shooting (YOUSAF AND HHB DATA)


(a) Pre-treatment Trends in Democratic vote share in Shooting Counties (HHB)

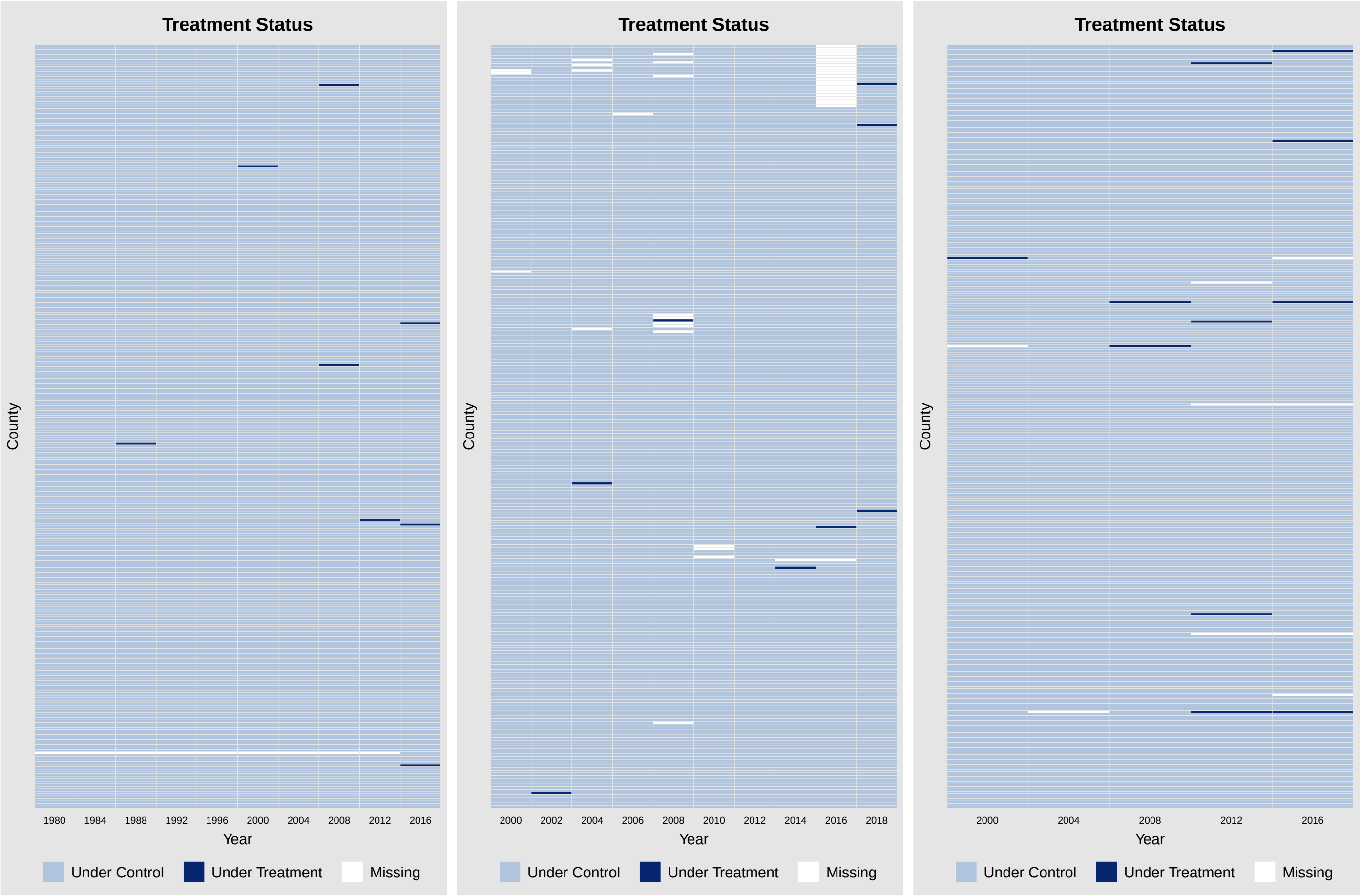
(b) Trends in Democratic vote share in Non Shooting Counties (HHB)

(c) Pre-treatment Trends in Democratic vote share in Shooting Counties (Yousaf)

(d) Trends in Democratic vote share in Non Shooting Counties (Yousaf)

Pre-treatment trends of Democratic vote share in counties where a shooting occurred and benchmarks this to the trends in Democratic vote share found in counties where a shooting did not occur for the Yousaf and HHB data. In the panels on the left, the small blue lines mark the patterns for all counties with a shooting and the bolded **blue** lines capture the average trend across these counties. The panels on the right show the same pattern for counties without a shooting. The small red lines mark the patterns for all counties without a shooting and the bolded **red** lines shows a loess model for counties without a shooting. **Takeaway:** Though taking a slightly different shape that the GMAL data, both the HHB and YOUSAF datasets show a separation between pre-treatment counties and control counties.

Figure S12: Treatment Across Counties Over Time, Only County Years with a Shooting are Treated



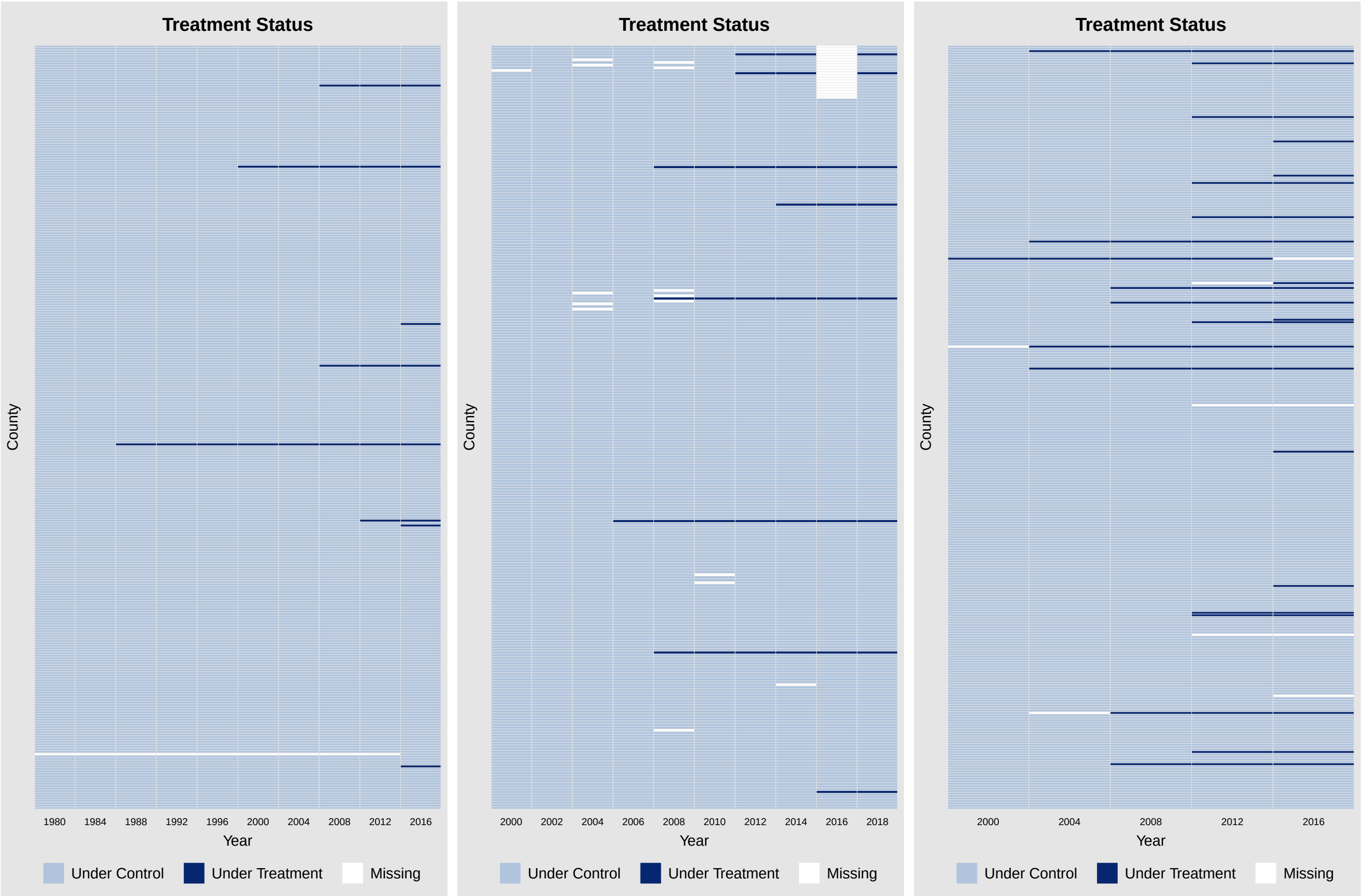
(a) GMAL Treatment Panel for Random Sample

(b) HHB Treatment Panel for Random Sample

(c) Yousaf Treatment Panel for Random Sample

Treatment over time for a random sample of counties in the three datasets illustrating treatment approach 1. Separate random counties are used in the figure that follows this one.

Figure S13: Treatment Across Counties Over Time, All Post Shooting Counties are Treated



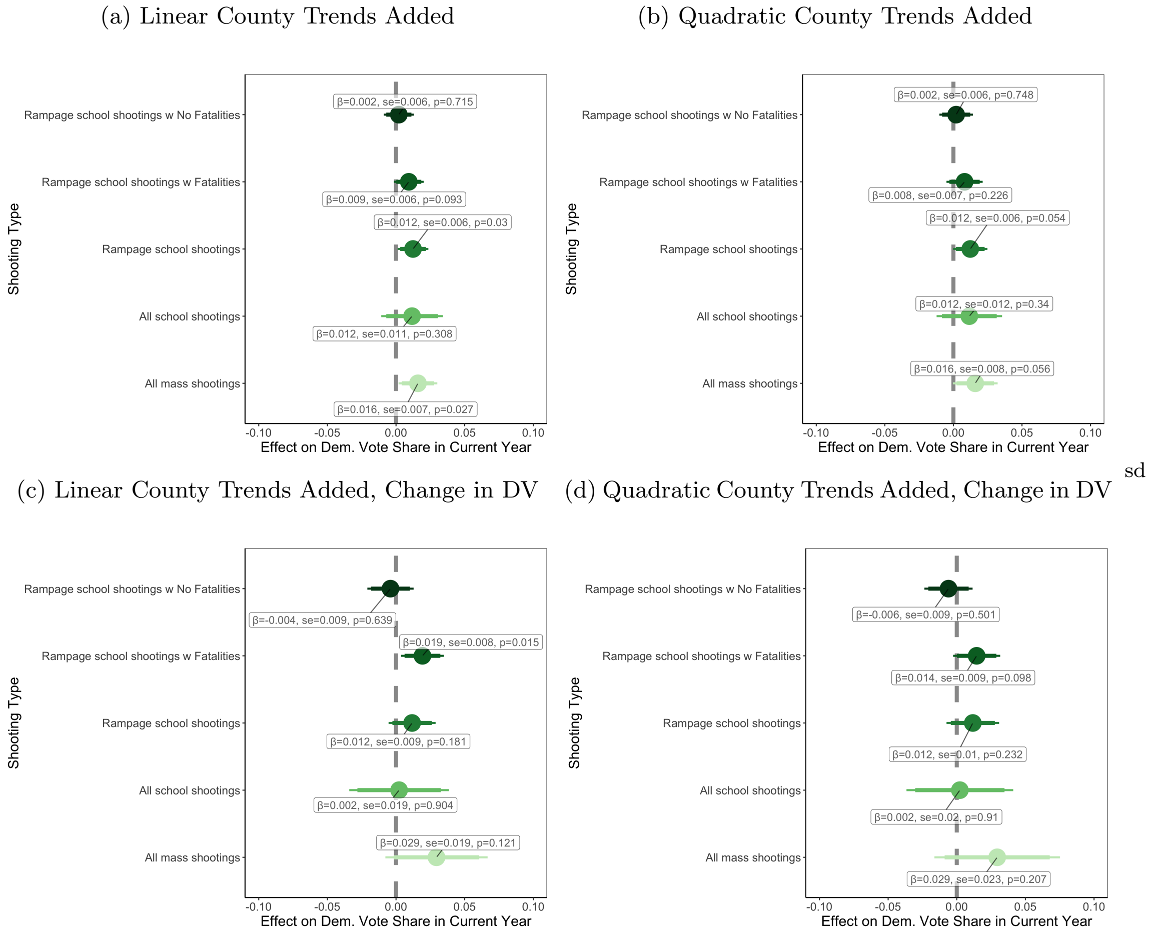
(a) GMAL Treatment Panel for Random Sample

(b) HHB Treatment Panel for Random Sample

(c) Yousaf Treatment Panel for Random Sample

Treatment over time for a random sample of counties in the three datasets illustrating treatment approach 2. Separate random counties are used in the figure that precedes this one.

Figure S14: The Effect of Mass Shootings on Presidential Election Returns Once County-Specific Trends are Absorbed, All Post Shooting Counties are Treated



(a) Linear County Trends Added

(b) Quadratic County Trends Added

sd

(c) Linear County Trends Added, Change in DV

(d) Quadratic County Trends Added, Change in DV

Effect of mass shootings of various types once we account for differential trends in Democratic vote share across counties in the United States. Within each panel, the first 3 estimates are using the GMAL coding of mass shootings and their data, the next comes from HHB, and the last comes from Yousaf. The upper left panel shows specifications with linear county trends, the upper right panel shows specifications with quadratic county trends, the bottom left panel shows specifications with linear county trends and using a change in Democratic vote share over the prior 4-year-previous election, the bottom right panel shows specifications with quadratic county trends and using a change in Democratic vote share over the prior 4-year-previous election. Coefficients, standard errors, and p-values are labeled for each coefficient. **Takeaway:** Once we account for differential trends across counties, the effects of mass shootings—be they located on school grounds or not, or be they rampage style or not—are all small and precisely-estimated.

Table S2: The ATT for each period, across all groups or cohorts (GMAL)

stats	Average	T1984	T1988	T1992	T1996	T2000	T2004	T2008	T2012	T2016
b	.0518245	.0619944	.0245811	.0246506	.0364387	.053292	.0565589	.0617882	.0599066	.0872101
se	.0096223	.0136641	.011968	.01148	.0120084	.0125972	.013094	.0141372	.0135578	.0118054
z	5.385884	4.537023	2.053896	2.147266	3.034426	4.230466	4.319438	4.370606	4.418599	7.387303
pvalue	7.21e-08	5.71e-06	.0399858	.0317721	.0024099	.0000233	.0000156	.0000124	9.93e-06	1.50e-13
ll	.0329652	.0352133	.0011242	.0021502	.0129026	.028602	.030895	.0340798	.0333337	.064072
ul	.0706839	.0887756	.0480381	.0471509	.0599748	.0779821	.0822227	.0894967	.0864794	.1103483

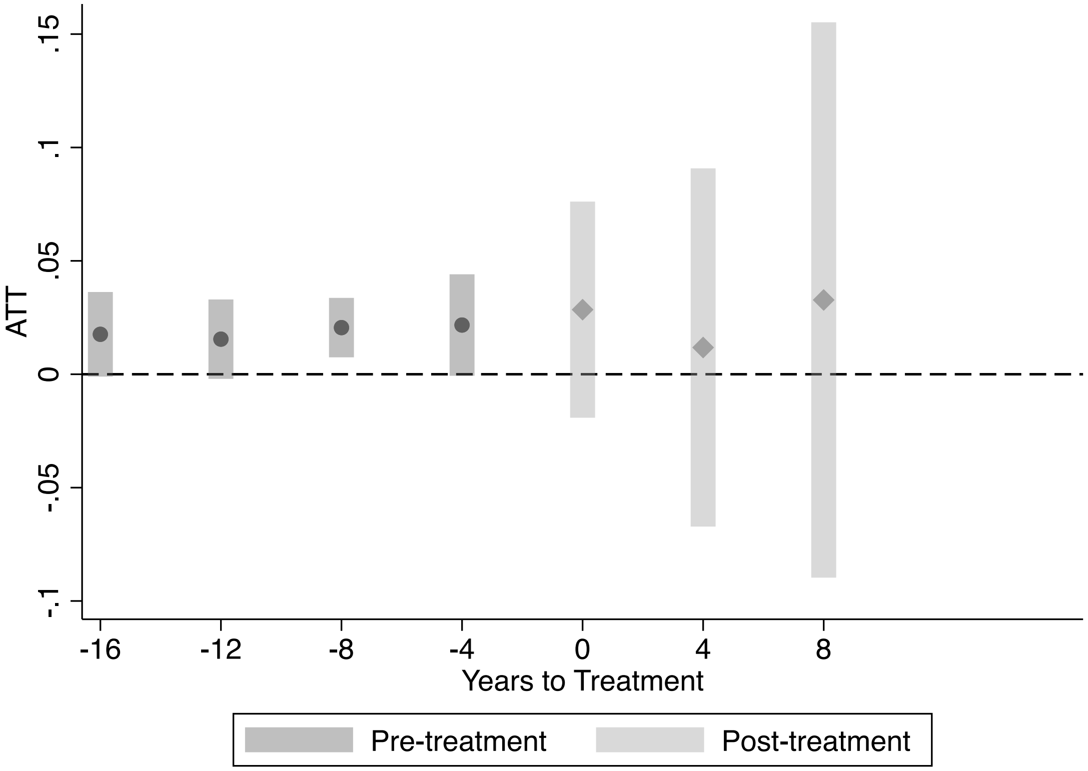
Estimates of the ATT for each period, across all groups or cohorts (i.e. the "Calendar" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Table S3: The ATT for each group or cohort, across all periods (GMAL)

stats	Average	G1984	G1988	G1992	G1996	G2000	G2004	G2008	G2012	G2016
b	.053623	.1587951	.0391324	.0485951	.0563038	.0283724	.0285435	.0474233	.0469285	.0565588
se	.0068437	.0189939	.0201407	.0162478	.0424463	.0300613	.0270402	.0116037	.006478	.0069854
z	7.835439	8.360334	1.942957	2.990872	1.326472	.9438179	1.055595	4.086901	7.244247	8.096711
pvalue	4.67e-15	6.25e-17	.0520213	.0027818	.1846836	.3452627	.2911533	.0000437	4.35e-13	5.65e-16
ll	.0402097	.1215678	-.0003425	.01675	-.0268894	-.0305467	-.0244543	.0246804	.0342318	.0428676
ul	.0670363	.1960224	.0786074	.0804402	.139497	.0872915	.0815413	.0701663	.0596252	.0702499

Estimates of the ATT for each group or cohort, across all periods (i.e. the "Group" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Figure S15: Estimation of all Dynamic Effects (GMAL)



Estimates of the dynamic effects (i.e. the “Event” estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant’Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** Pre-treatment imbalances can be seen in the figure. This suggests that *even when* one uses “clean comparions” as suggested by Callaway and Sant’Anna (2021), differential pre-treatment trends are an issue. At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice. We reference the reader to the event study estimates in the paper for those that adjust for differential trends identified in the paper

Table S4: The ATT for each period, across all groups or cohorts (HHB)

stats	Average	T2002	T2004	T2006	T2008	T2010	T2012	T2014	T2016	T2018
b	.032836	.051527	.0243792	-.0458838	.0395123	.0239998	.0489557	.0240271	.0796322	.0493749
se	.0229896	.0362721	.0228559	.0475194	.0432784	.0368888	.0285933	.0246188	.0207751	.0173127
z	1.4283	1.420568	1.066648	-.9655811	.9129784	.650598	1.712138	.9759645	3.833059	2.851938
pvalue	.1532057	.1554425	.2861309	.3342539	.3612539	.515306	.0868712	.3290821	.0001266	.0043454
ll	-.0122227	-.0195651	-.0204175	-.1390202	-.0453119	-.0483009	-.0070862	-.0242249	.0389137	.0154425
ul	.0778948	.1226191	.0691759	.0472525	.1243364	.0963005	.1049976	.072279	.1203506	.0833072

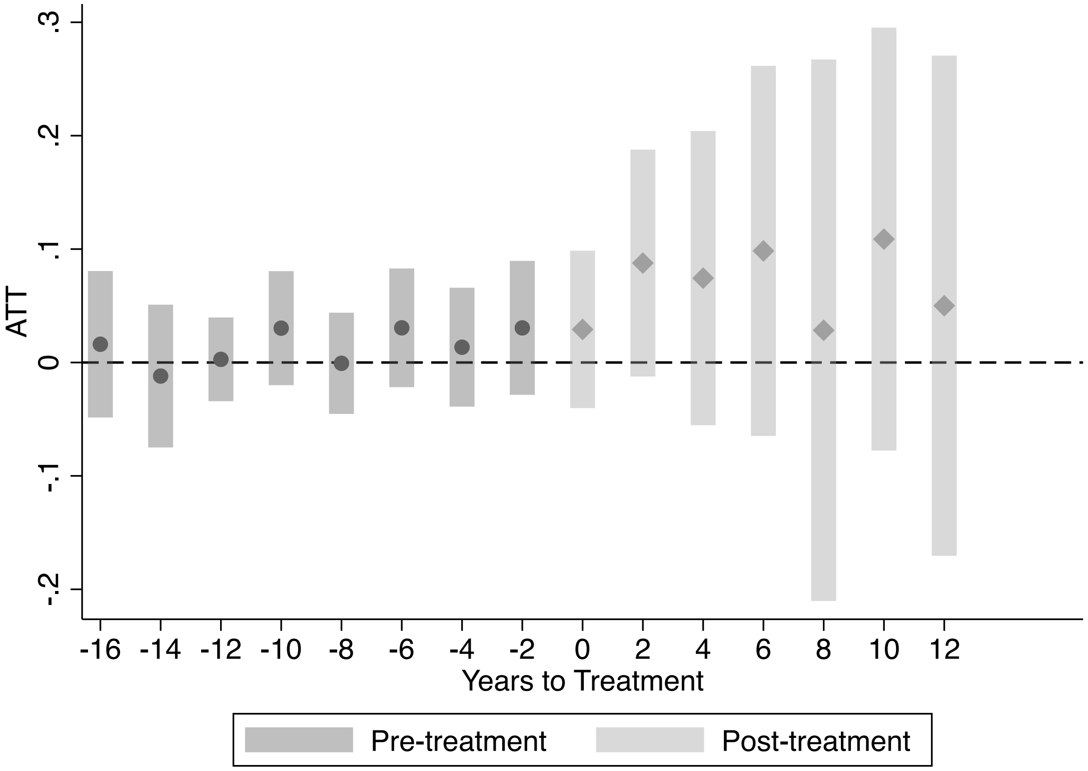
Estimates of the ATT for each period, across all groups or cohorts (i.e. the “Calendar” estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant’Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Table S5: The ATT for each group or cohort, across all periods (HHB)

states	Average	G2002	G2004	G2006	G2008	G2010	G2012	G2014	G2016	G2018
b	.0323525	.0222296	.056052	-.0032929	.1754662	.137212	.0905926	.010359	.000745	.0107809
se	.0142005	.0621462	.0256692	.0714713	.0603517	.0051282	.0259543	.0148958	.0178077	.0251306
z	2.278258	.3576986	2.183627	-.0460732	2.907397	26.75635	3.490473	.6954327	.0418356	.4289939
pvalue	.0227112	.7205689	.0289897	.9632519	.0036445	1.0e-157	.0004822	.4867841	.9666297	.6679277
ll	.0045199	-.0995748	.0057412	-.1433741	.0571791	.1271609	.0397232	-.0188362	-.0341574	-.0384742
ul	.060185	.144034	.1063627	.1367882	.2937533	.1472631	.141462	.0395542	.0356474	.0600359

Estimates of the ATT for each group or cohort, across all periods (i.e. the “Group” estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant’Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Figure S16: Estimation of all Dynamic Effects (HHB)



Estimates of the dynamic effects (i.e. the “Event” estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant’Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** Pretreatment imblances are of least concern in the HHB data, and this is where we observe no evidence for a significant effect. At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice. We reference the reader to the event study estimates in the paper for those that adjust for differential trends identified in the paper

Table S6: The ATT for each period, across all groups or cohorts (Yousaf)

stats	Average	T2004	T2008	T2012	T2016
b	.0387695	.015784	.0336756	.035597	.0700214
se	.0088303	.0117028	.0116784	.0173425	.0095523
z	4.390519	1.348739	2.88359	2.052591	7.330298
pvalue	.0000113	.1774209	.0039317	.0401123	2.30e-13
ll	.0214625	-.007153	.0107865	.0016064	.0512992
ul	.0560765	.0387209	.0565648	.0695876	.0887436

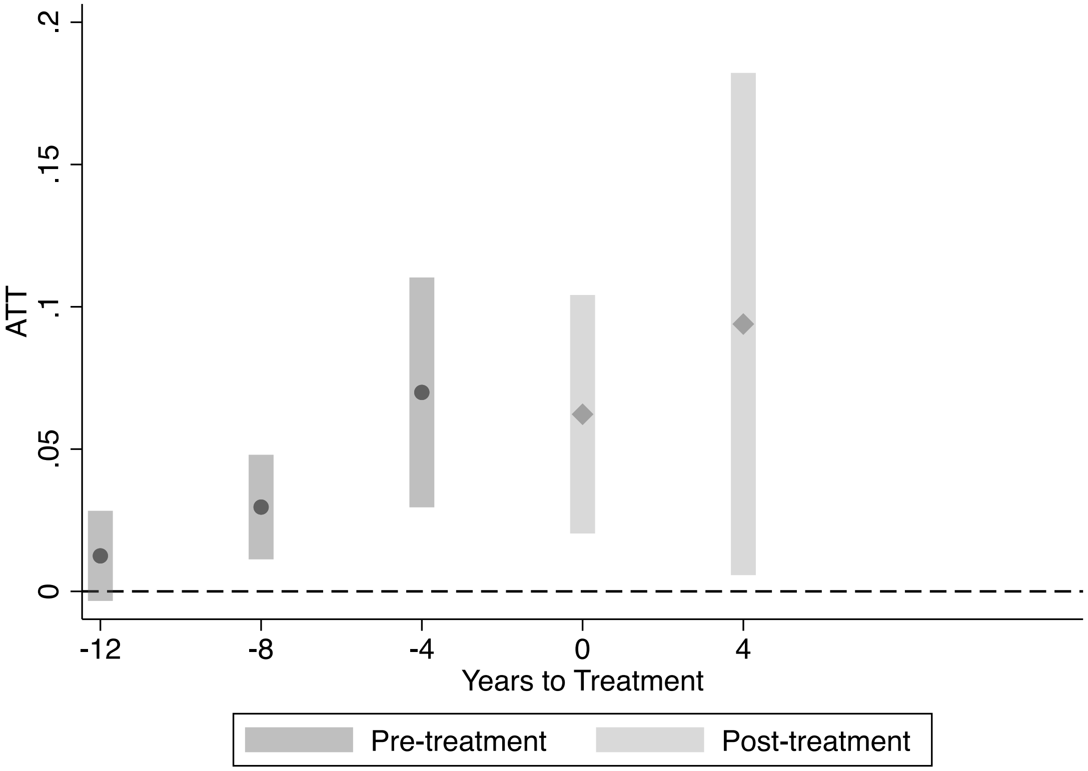
Estimates of the ATT for each period, across all groups or cohorts (i.e. the "Calendar" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Table S7: The ATT for each group or cohort, across all periods (Yousaf)

stats	Average	G2004	G2008	G2012	G2016
b	.0496422	.0476684	.0615136	.0460837	.0477795
se	.0071074	.0205126	.0151624	.0098212	.0115534
z	6.984585	2.323858	4.056979	4.692281	4.135519
pvalue	2.86e-12	.0201331	.0000497	2.70e-06	.0000354
ll	.035712	.0074644	.0317958	.0268345	.0251351
ul	.0635725	.0878725	.0912314	.0653328	.0704238

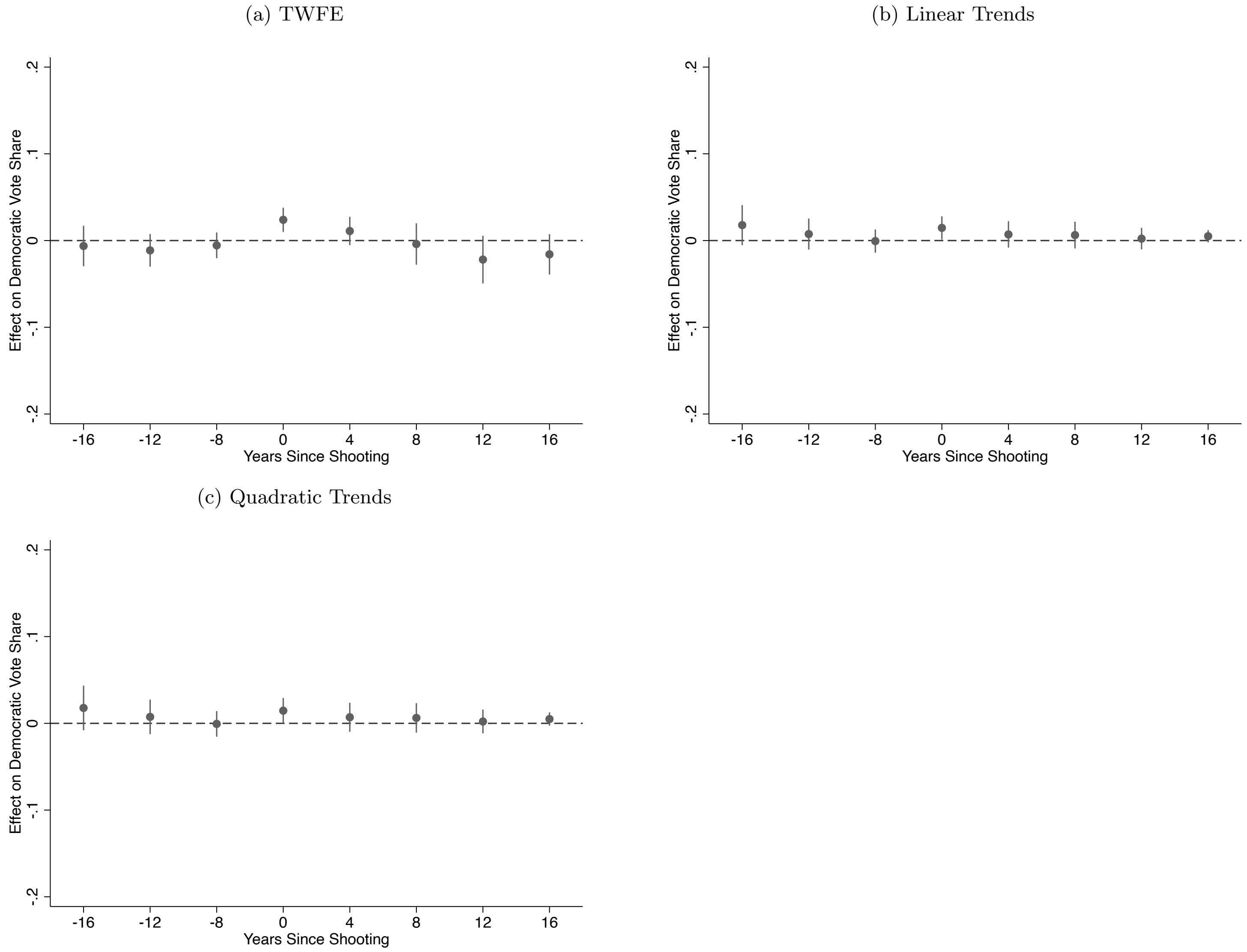
Estimates of the ATT for each group or cohort, across all periods (i.e. the "Group" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice.

Figure S17: Estimation of all Dynamic Effects (Yousaf)



Estimates of the dynamic effects (i.e. the "Event" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. **Takeaway:** Pre-treatment imbalances can be seen in the figure. This suggests that *even when* one uses "clean comparions" as suggested by Callaway and Sant'Anna (2021), differential pre-treatment trends are an issue. At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice. We reference the reader to the event study estimates in the paper for those that adjust for differential trends identified in the paper

Figure S18: Sun and Abraham (2020) Event Study Estimates (GMAL)



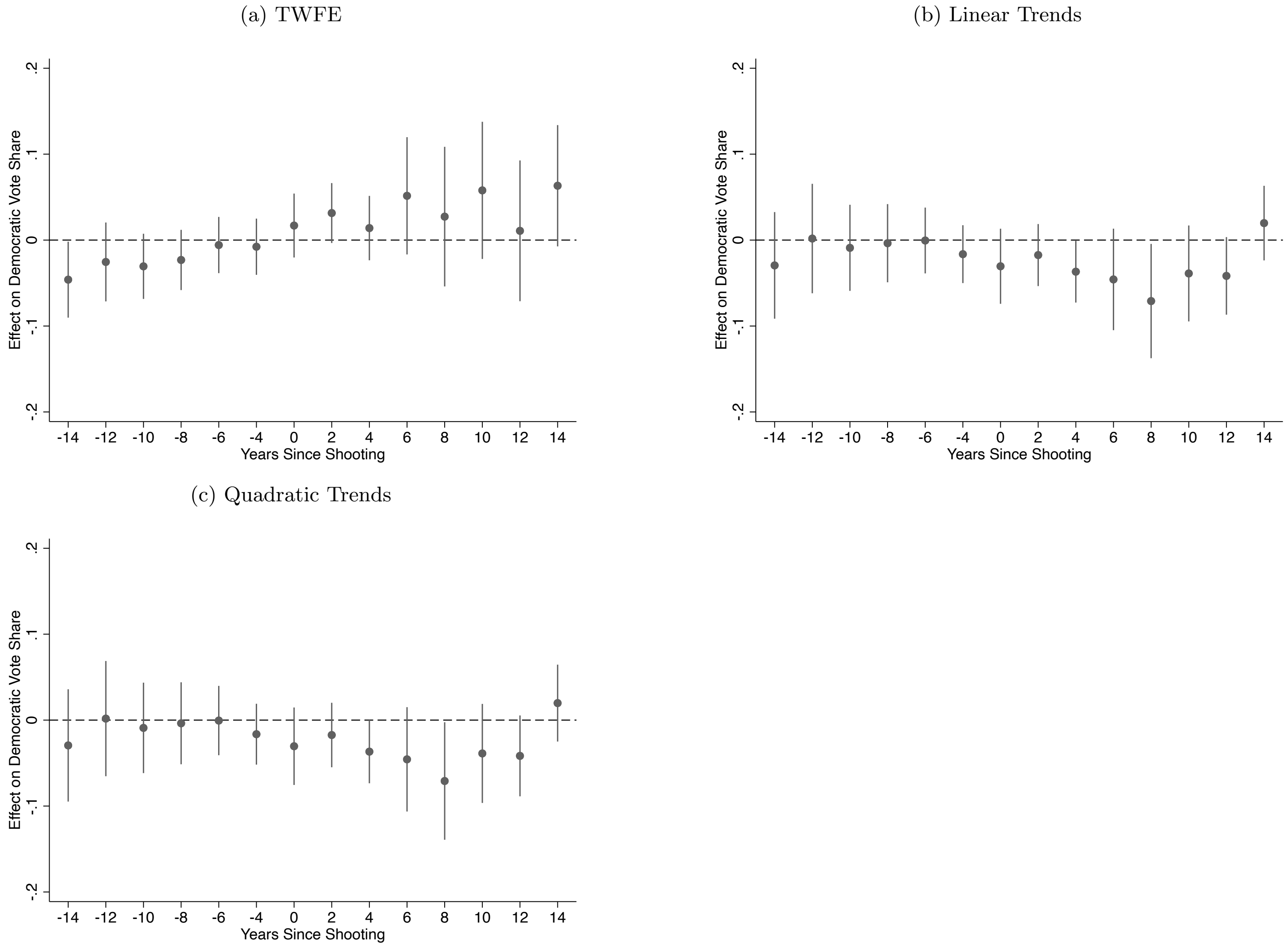
(a) TWFE

(b) Linear Trends

(c) Quadratic Trends

Sun and Abraham (2020) event study estimates through the `eventstudyinteract` package provided by the authors. Standard errors are clustered at the county level. **Takeaway:** Clean comparison effects with trends show no sign of a sizable and durable effect on Democratic vote shares shown in the TWFE nor in the simple event study plot

Figure S19: Sun and Abraham (2020) Event Study Estimates (HHB)



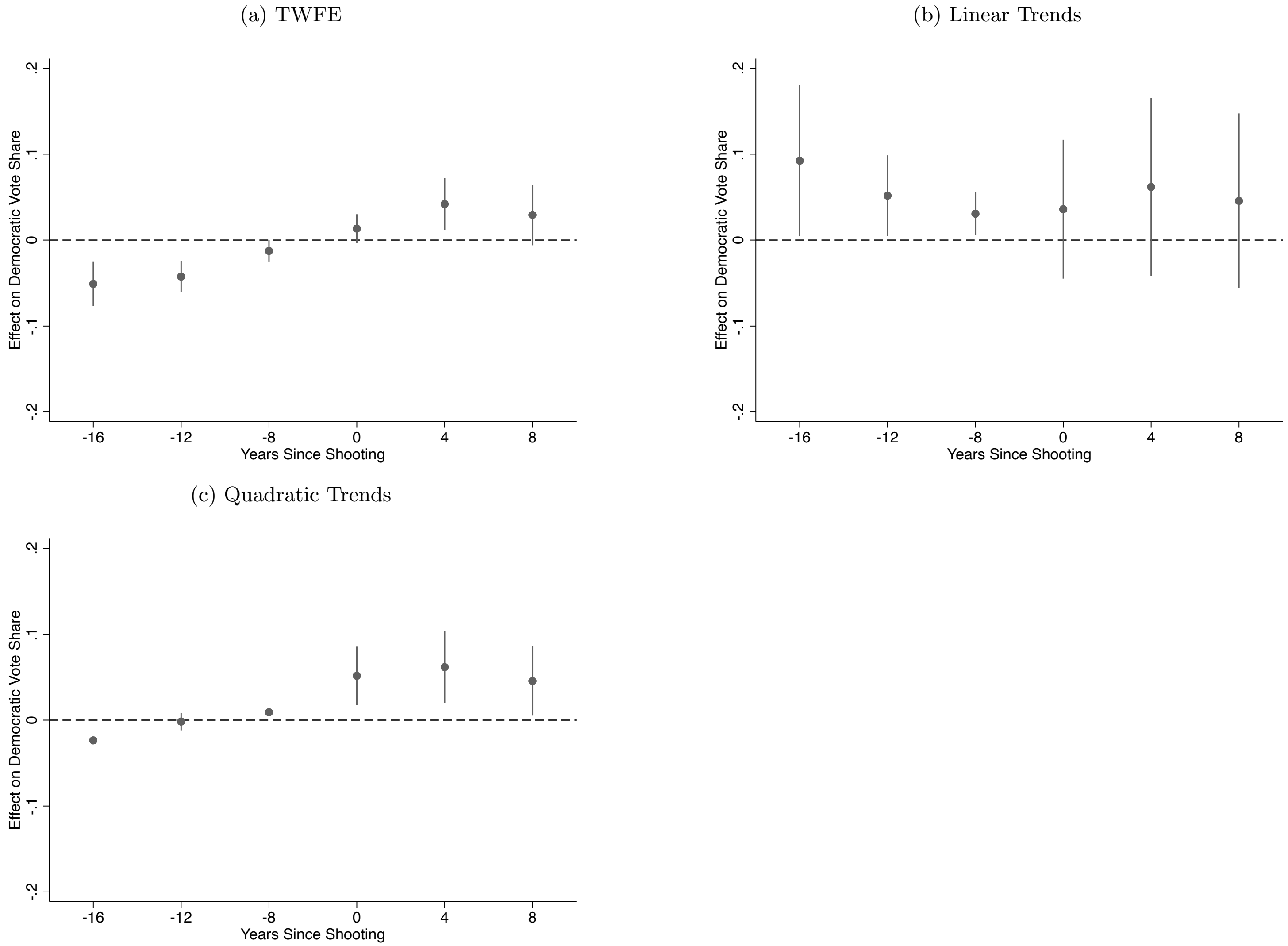
(a) TWFE

(b) Linear Trends

(c) Quadratic Trends

Sun and Abraham (2020) event study estimates through the `eventstudyinteract` package provided by the authors. Standard errors are clustered at the county level. **Takeaway:** Clean comparison effects with trends show no sign of a sizable and durable effect on Democratic vote shares shown in the TWFE nor in the simple event study plot

Figure S20: Sun and Abraham (2020) Event Study Estimates (Yousaf)



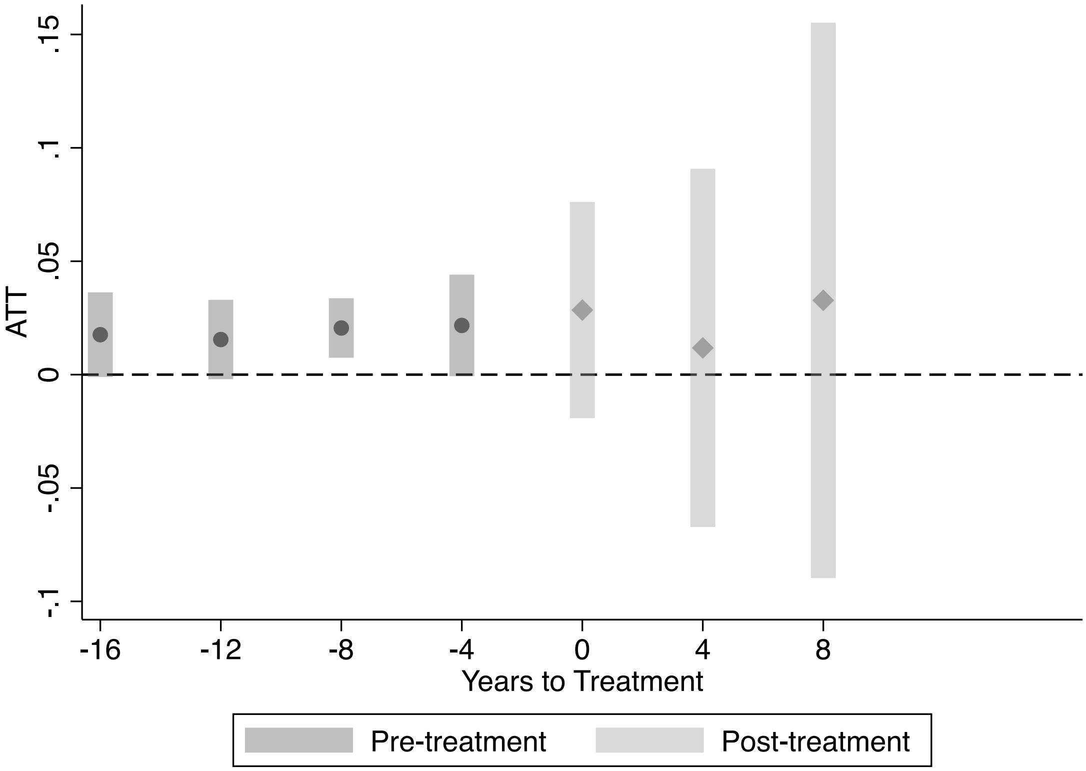
(a) TWFE

(b) Linear Trends

(c) Quadratic Trends

Sun and Abraham (2020) event study estimates through the `eventstudyinteract` package provided by the authors. Standard errors are clustered at the county level. **Takeaway:** Clean comparison effects with trends show no robust sign of a sizable and durable effect on Democratic vote shares shown in the TWFE nor in the simple event study plot. The trends specification for this approach in the Yousaf data still show signs of pre-treatment imbalance and, as such, should be interpreted with care.

Figure S21: Estimation of Clean Comparison TWFE Effects using the Callaway and Sant'Anna (2021) Approach



Estimates of the dynamic effects (i.e. the "Event" estimates provided in the CSdid package) based on the procedure developed by Callaway and Sant'Anna (2021). Estimates use the doubly Robust IPW (DRIPW) estimation method, with Wildbootstrap SE, and not-yet treated observations as controls. Controls used by GMAL are included—i.e. population, proportion non-white, and change in the unemployment rate. **Takeaway:** Pre-treatment imbalances can still be seen in the figure. This suggests that *even when* one uses "clean comparions" as suggested by Callaway and Sant'Anna (2021) and covariates, differential pre-treatment trends are an issue. At present, this method does not allow for the inclusion of unit-present trends, so estimates for this empirical case should be viewed with caution. These are included as an illustration of how to use this method in practice. We reference the reader to the event study estimates in the paper for those that adjust for differential trends identified in the paper.

Table S8: Estimation of Clean Comparison TWFE Effects using the de Chaisemartin and D'Haultfoeuille Approach – HHB

	Estimate	SE	LB CI	UB CI	N	Switchers
Effect_0	0.0124381	0.0186208	-0.0240587	0.048935	26791	91
Effect_1	0.0180224	0.0268722	-0.0346472	0.0706919	23526	71
Effect_2	0.0051758	0.0350417	-0.063506	0.0738576	20319	61
Effect_3	0.0493664	0.0762519	-0.1000873	0.1988201	17146	34
Effect_4	0.0116177	0.1171154	-0.2179286	0.2411639	14088	26
Placebo_1	0.016137	0.0113333	-0.0060763	0.0383503	23540	85
Placebo_2	-0.0214889	0.0152387	-0.0513567	0.0083788	20341	83
Placebo_3	0.0211926	0.0126292	-0.0035607	0.0459458	17184	72
Placebo_4	-0.0103004	0.0117131	-0.0332582	0.0126574	14126	64

Note: de Chaisemartin and D'Haultfoeuille (2020) approach for assessing and addressing implemented in the `did_multiplegt` package in *STATA* and `DIDmultiplegt` package in *R*. Under the common trends assumption, beta estimates a weighted sum of 395 ATTs. 379 ATTs receive a positive weight, and 16 receive a negative weight. The sum of the positive weights is equal to 1.0010116. The sum of the negative weights is equal to -.00101162. beta is compatible with a DGP where the average of those ATTs is equal to 0, while their standard deviation is equal to .12133344. beta is compatible with a DGP where those ATTs all are of a different sign than beta, while their standard deviation is equal to 13.249181. **Takeaway:** After using this method, we see no evidence of substantial or significant effects of mass shootings on electoral outcomes.

Table S9: Estimation of Clean Comparison TWFE Effects using the de Chaisemartin and D'Haultfoeuille Approach – GMAL

	Estimate	SE	LB CI	UB CI	N	Switchers
Effect_0	0.0170459	0.004788	0.0076614	0.0264304	27632	98
Effect_1	0.0159759	0.0101364	-0.0038914	0.0358431	24507	72
Effect_2	0.0227205	0.0188645	-0.0142539	0.0596949	21409	59
Effect_3	0.0359566	0.0292597	-0.0213925	0.0933056	18315	47
Effect_4	0.0797536	0.0415763	-0.0017359	0.1612432	15231	38
Placebo_1	0.0000278	0.004569	-0.0089274	0.0089831	24526	91
Placebo_2	0.006823	0.0051979	-0.0033649	0.0170109	21429	79
Placebo_3	-0.0031506	0.0038482	-0.010693	0.0043917	18344	76

Note: de Chaisemartin and D'Haultfoeuille (2020) approach for assessing and addressing implemented in the `did_multiplegt` package in *STATA* and `DIDmultiplegt` package in *R*. Under the common trends assumption, beta estimates a weighted sum of 400 ATTs. 396 ATTs receive a positive weight, and 4 receive a negative weight. The sum of the positive weights is equal to 1.0002424. The sum of the negative weights is equal to -.00024243. beta is compatible with a DGP where the average of those ATTs is equal to 0, while their standard deviation is equal to .13523432. beta is compatible with a DGP where those ATTs all are of a different sign than beta, while their standard deviation is equal to 30.161087. **Takeaway:** After using this method, we see no evidence of substantial or significant effects of mass shootings on electoral outcomes. Effect_0 is not robust to other approaches for adjusting for potential violations of the parallel trends assumption—e.g. Rambachan and Roth (2021).

Given Figure 2 in the manuscript, some may wonder if we discard never treated units and, instead, compare the treated units with the not-yet-but-eventually-treated units. Such could be valid comparison group. If they were, we could, perhaps, avoid taking a stand on the type of violations of parallel trends. Unfortunately, this is not the case. We still observe pre-treatmenet imbalances among this group. These are of similar magnitude to the effects observed post-treatment. Once trends are added, any evidence for an effect disappears. This is shown in the Table below. Though this approach doesn't work in ours, this comparison could be a viable option for applied researchers in other settings.

Table S10: Using Eventually Treated as the Control Group

treatment	time	YearsPre	model	coef	tstat	stderr	pval	N	r2
All school shootings	Post	-4	Quad Trends	.006	1.378	.004	.171	990	.946
All school shootings	Post	-4	Linear Trends	.006	1.471	.004	.144	990	.946
All school shootings	Post	-4	TWFE	.018	2.452	.007	.016	990	.794
All school shootings	Pre	20	TWFE	.015	2.446	.006	.016	495	.886
All school shootings	Pre	16	TWFE	.009	1.585	.005	.116	594	.875
All school shootings	Pre	12	TWFE	.006	.992	.006	.324	693	.862
All school shootings	Pre	8	TWFE	.019	2.728	.007	.008	792	.837
All school shootings	Pre	4	TWFE	.01	1.539	.007	.127	891	.815

Table S11: Adding Year Trends to Covariates in GMAL Data

model	coef	stderr	tstat	pval	N	r2
quad county trends with covs linear trends	.002	.005	.53	.596	18620	.968
linear county trends with covs linear trends	.003	.004	.599	.549	18620	.968
quad trends with covs controlled	.007	.005	1.504	.133	18620	.968
linear trends with covs controlled	.007	.004	1.684	.092	18620	.968
quad trends with no covs omit missing covs	.007	.005	1.578	.115	18620	.967