Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

An Evaluation of Difference-in-Differences Methods Using Placebo Event Studies

John Coglianese and Jade A. Fang

2026-045

Please cite this paper as:
Coglianese, John, and Jade A. Fang (2026). "An Evaluation of Difference-in-Differences Methods Using Placebo Event Studies," Finance and Economics Discussion Series 2026-045. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2026.045.

An Evaluation of Difference-in-Differences Methods Using Placebo Event Studies

John Coglianese ⓡ Jade A. Fang *

Board of Governors of the Federal Reserve System

June 1, 2026

Abstract

Researchers are faced with the choice of which of the many recently developed difference-in-differences methods to use in practice. To assess these estimators' relative performance for single-unit event studies, we conduct 134,000+ state-level placebo event studies across 13 estimators. We find that no single method dominates. Performance is context-dependent, with synthetic-control-like methods sometimes outperforming and sometimes underperforming two-way-fixed-effect-like and matching methods. Performance also varies across states at least as much as it does across estimators. Our results highlight the need for practitioners to conduct placebo tests to understand the performance of methods in their research context.

*The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the Board of Governors of the Federal Reserve System or its staff. The circled r between authors denotes that author order was randomized following https://www.aeaweb.org/journals/random-author-order.

1 Introduction

Difference-in-differences is a popular research design for estimating the causal effects of policies, events, and other local economic shocks. Traditionally, practitioners have estimated difference-in-differences using two-way fixed effects regressions (TWFE); however, a growing literature has documented problems with TWFE in staggered adoption settings and offered alternative estimators for difference-in-differences (e.g., Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Roth et al., 2023). Researchers using difference-in-differences are now faced with many choices in practice, including which estimator to use, whether to include controls, and whether to normalize.

In this paper, we use placebo tests to evaluate the empirical performance of difference-in-differences estimators. For each estimator, we conduct event studies repeatedly over thousands of randomly chosen "events," which we then aggregate to assess estimator efficiency based on the spread of placebo estimates. Our results suggest that there is no single "superior" method. Synthetic-control-like methods sometimes outperform and sometimes underperform TWFE-like methods, depending on the estimator and the outcome variable. Other choices also impact performance at least as much as the estimator, including seasonal adjustment, normalization, and which units are "treated." Given such context dependence, we urge practitioners to conduct placebo tests to understand the performance of their methods in their unique setting.

Placebo tests are advantageous in evaluating difference-in-differences methods. Placebo tests are a form of randomization inference, as the random sampling of placebo events reveals the distribution of event study estimates underlying the null hypothesis. In this way, placebo tests measure the efficiency of estimators through the spread of the distribution of placebo estimates. Furthermore, placebo tests use real-world data and thus circumvent making strict assumptions about the data generating process, which Monte Carlo simulations require. Ultimately, we view placebo tests as complementary to econometric theory and simulation results in helping practitioners decide which methodology to use in their research design.

At a high level, we conduct placebo tests by randomly drawing some "event" and estimating an event study as if some real treatment were applied in that event. For example, we draw an "event" in Nebraska for October 1998 and then implement a difference-in-differences estimator (e.g., TWFE, synthetic control) to obtain event study estimates comparing Nebraska to control states over a range spanning before and after October 1998. Repeating this process with additional randomly sampled "events" produces a distribution of placebo estimates for the estimator and outcome of interest.

From these distributions of placebo estimates, we can compare difference-in-differences methods in terms of performance. A distribution centered at zero indicates no bias, and a lower variance indicates better efficiency. We consider three families of estimator: TWFE-like estimators, which include variations of TWFE designed to avoid staggered adoption issues (e.g., Callaway and Sant'Anna, 2021); matching estimators, which subset control units based on covariates (Rosenbaum and Rubin, 1983; Iacus et al., 2012); and synthetic-control-like estimators, which construct counterfactuals for the treatment unit flexibly based on pre-treatment data (e.g., Abadie et al., 2010; Arkhangelsky et al., 2021). In our placebo tests, we use common outcomes studied in research settings, including the unemployment rate, payroll employment, and a house price index, all measured at the state level in the United States. Our main comparison across methods involves calculating the mean squared error of each estimator's placebo distribution and normalizing the percent difference relative to TWFE, which gives the ordering of estimators by efficiency.

Our results indicate that there is no single "superior" estimator. TWFE-like estimators and matching estimators perform relatively similarly to TWFE in our context.[1] However, synthetic-control-like estimators are much more variable across estimators and outcomes and in different cases can either outperform or underperform TWFE by significant margins. We find no systematic pattern in the ranking of estimators by efficiency across outcomes.

We interpret the variation in performance of synthetic-control-like estimators as evi-

[1]Note that we focus on the single event study setting in our analysis; conducting placebo tests in staggered adoption settings may result in more discernable differences in performance for the TWFE-like estimators.

dence of a bias-variance trade-off for difference-in-differences. Methods that optimize on pre-treatment outcomes, as synthetic-control-like methods commonly do, can sometimes construct better counterfactuals by picking up on heterogeneous dynamics across control units. However, the flexibility of these models in forming counterfactuals can also lead to overfitting in other cases. Practitioners should be aware when using these methods that their performance is context-dependent.

We also find that certain choices aside from the estimator can affect performance. On average, normalization to the last pre-treatment period and seasonal adjustment of the outcome variable meaningfully lower the variance of placebo estimates, while the choice of covariates (for estimators that use them) matters considerably less.

Lastly, we find that which *unit* is treated impacts performance at least as much as which *estimator* is chosen. In our state-level analysis, we identify a pattern that persists across combinations of estimators and outcomes: placebo events in large, diverse states (e.g., Texas, Pennsylvania, Ohio) systematically yield lower variance of placebo estimates than their idiosyncratic counterparts (e.g., Nevada, Alaska, Hawaii). We conjecture that large, diverse states better resemble the United States as a whole, so methods can more easily approximate outcomes for these states from the set of remaining control states.

Our paper adds to the growing literature on difference-in-differences methods (Abadie et al., 2010; Xu, 2017; de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Arkhangelsky et al., 2021; Athey et al., 2021; Roth et al., 2023; Borusyak et al., 2024; Dube et al., 2025; Baker et al., 2025). While this literature largely proposes and evaluates new estimators based on econometric theory, we shed light on the relative performance of these estimators in practice.

Our results from placebo tests also complement the literature evaluating difference-in-differences methods with Monte Carlo simulations (Bertrand et al., 2004; Arkhangelsky et al., 2021; Borusyak et al., 2024). Monte Carlo simulations have advantages in that they allow studying estimator performance precisely under different data generating processes.

However, the results may be specific to the data generating process used, and they may or may not be relevant to real-world settings depending on whether the assumed data generating process resembles real-world data. Our placebo tests provide a direct test of estimator performance that researchers may face in practice.

While econometric theory and Monte Carlo simulations are the main methods used in prior literature, some earlier analyses have used placebo tests to evaluate difference-in-differences (e.g., Bertrand et al., 2004). We build on these earlier placebo tests by examining a comprehensive set of difference-in-differences methods, including recently-developed methods. We additionally use placebo tests to highlight the importance of research setting, which matters in practice at least as much as estimator choice.

Our results will be useful for practitioners conducting local labor market event studies. This type of analysis includes a large and growing share of papers in applied microeconomics (Card and Krueger, 1994; Neumark and Wascher, 2000; Autor et al., 2013; Chodorow-Reich et al., 2019; Cengiz et al., 2019; Hornbeck and Moretti, 2024). These papers often employ difference-in-differences for causal inference, so our evaluation of commonly used methods contributes novel insights that may help researchers choose which estimator to implement and how.

The rest of the paper is organized as follows. Section 2 justifies our use of placebo tests for evaluating estimator performance. Section 3 explains our placebo testing approach, provides an overview of our 13 difference-in-differences methods of interest, and summarizes the technical details of our methodological implementation. Section 4 reports the relative performance of estimators based on placebo event studies. Section 5 highlights practitioner's choices conditional on estimator that may or may not improve precision. Section 6 emphasizes the significance of the research setting in shaping performance. Finally, Section 7 concludes and offers advice for practitioners.

2 Why placebo tests?

Our placebo testing approach is informative in evaluating difference-in-differences methods because it reveals the distribution of estimates underlying the null hypothesis, essentially a form of randomization inference (Fisher, 1935). By randomizing placebo events in each state and year and then aggregating results by outcome and estimator, we can estimate the variance of difference-in-differences estimates, assuming an "effect" of zero.[2] A narrower distribution of estimates implies a lower standard error under randomization inference, which is indicative of a more efficient estimator for the outcome of interest.

Placebo tests are complementary to Monte Carlo simulations. Both methodologies use randomization to reveal the finite-sample properties of estimators (Bertrand et al., 2004; Arkhangelsky et al., 2021; Borusyak et al., 2024). However, the two differ in their approaches: placebo tests leverage real-world data, while Monte Carlo simulations assume specific data generating processes. Monte Carlo simulations are advantageous because they give the analyst full control over the data generating process, but the results are specific to said data generating process and may not necessarily be representative of real-life dynamics. By incorporating real-world data, placebo tests offer the analyst a unique opportunity to conduct more realistic assessments of estimator performance.

Through placebo tests, we can also understand estimator performance in an empirical setting commonly used by researchers (Bertrand et al., 2004; Arkhangelsky et al., 2021). Econometric theory covers all possible settings, which provides broad insights into how different estimators may perform under various assumptions; in contrast, our placebo tests can compare estimator performance in only a specific setting: state-level labor market event studies in the United States. Nonetheless, understanding performance in this particular case may be more useful to applied microeconomists working in this setting than understanding

[2]As in usual randomization inference, this exercise is informative about the variance under the null hypothesis only for studying policies that do not change during our sample period. For policies that change during our sample period, our variance estimate also potentially includes the treatment effect, but one could conduct a similar exercise excluding the treated units to estimate the variance under the null hypothesis.

broader differences in performance.

Lastly, placebo tests help to emphasize the role of setting in econometric performance, beyond differences between estimators. Some states may be more idiosyncratic due to demographic or economic factors, while other states will be more similar to the national average. Such differences could lead to lower efficiency for difference-in-differences analyses conducted in the more idiosyncratic states, which placebo tests reveal through more dispersed estimates. In this way, placebo tests provide useful information about economic geography that practitioners can use when thinking about difference-in-differences settings.

3 Methodology

In the difference-in-differences research design, a placebo test involves estimating an event study with real data for randomly drawn "events." In the usual implementation of difference-in-differences event studies, events represent periods/units in which a policy changes, a shock occurs, or some other form of treatment is applied to the unit. Under the placebo testing framework, our events are "fake": since they are randomly drawn from the set of all possible events, we expect null effects on the outcome on average.

At a high level, each step of the placebo test involves randomly drawing an "event" and then estimating an event study as if some treatment truly occurred during that event. For example, we draw an "event" in Nebraska for October 1998. We then use a difference-in-differences estimator (two-way fixed effects, synthetic control, etc.) to obtain event study estimates comparing Nebraska to control states over a range spanning before and after October 1998. We then repeat this process many times with additional randomly drawn events and pool the estimates together to produce a distribution of placebo estimates, which is informative of the estimator's performance under the null hypothesis.

We produce such distributions for each estimator and outcome separately. For each distribution, the mean of placebo estimates represents the probability limit of the estimator

and the spread reflects its efficiency. Since we have randomly chosen our "events," all estimators should have the same probability limit of zero; however, estimators may vary in terms of their efficiency. Lower variance of placebo estimates indicates better performance in practice, which can guide researchers in choosing which estimator to implement in their research setting.

We aimed for a comprehensive overview of available difference-in-differences methods, selecting 13 methods used in recent literature. Table 1 lists our estimators of interest. We choose two-way fixed effects (TWFE) as the baseline and assign each of the other twelve estimators to one of the following groups: TWFE-like, matching, and synthetic-control-like. In the ensuing sections, we give a brief overview of each method, grouped by category.

Table 1: Difference-in-differences methods of interest

	Covariates	Optimize on pre-treatment outcomes	Normalization
Baseline			
Two-way fixed effects	n.a.	n.a.	Yes
TWFE-like			
Callaway and Sant'Anna	No	n.a.	Yes
Imputation	n.a.	n.a.	Yes
Wooldridge difference-in-differences	No	n.a.	Yes
De Chaisemartin & D'Haultfoeuille	n.a.	n.a.	Yes
Local projections	No	n.a.	Yes
Matching			
Propensity score matching	Yes	n.a.	Yes
Coarsened exact matching	Yes	n.a.	Yes
Synthetic-control-like			
Synthetic control	No	Yes	Yes
Elastic net synthetic control	n.a.	Yes	No
Generalized synthetic control	n.a.	Yes	No
Matrix completion	n.a.	Yes	No
Synthetic difference-in-differences	n.a.	Yes	No

Note: The table describes our implementation of each estimator in our baseline specification. n.a. means not applicable.

3.1 Two-way fixed effects

We focus on the setting of a single-unit event study: given a panel dataset with multiple units i and time periods t, exactly one unit receives treatment at a specific point in time. We can estimate an event study in this setting by implementing **two-way fixed effects (TWFE)**:

$$y_{i,t} = \alpha_i + \gamma_t + \sum_{\tau=-T}^{T} \beta^\tau \delta_{i,t-\tau} + \epsilon_{i,t}$$

In the equation above, our notation is as follows: $y_{i,t}$ is the outcome variable for unit i in time period t; α_i represents unit-fixed effects; and γ_t represents time-fixed effects. For each time horizon τ relative to the event, β^τ is the coefficient of interest, reflecting the dynamic treatment effect at τ periods after the event. The event dummies $\delta_{i,t-\tau}$ are leads and lags of the event occurrence for unit i at τ periods before/after time period t.

As the traditional method for estimating event studies, TWFE serves as our baseline for evaluating the relative performance of estimators. However, it implicitly makes comparisons across all units and time periods available in the data, which can sometimes include so-called "forbidden comparisons" across different cohorts of treated units. Accordingly, it suffers from well-documented problems in settings with staggered adoption and heterogeneous effects (e.g., Roth et al. (2023)). Econometricians have developed several novel estimators to mitigate such issues, which we examine in the next section.

3.2 TWFE-like estimators

We define TWFE-like estimators as variations of TWFE that are robust to the challenges of staggered adoption settings. In our single-event research setting, these distinctions are irrelevant: TWFE-like estimators all converge to the canonical TWFE estimator since we only have a single treated unit. However, these methods are slightly different in their implementation, so we still evaluate these methods to determine whether they perform any worse than TWFE under finite-sample randomization inference. We include in our analysis five

TWFE-like estimators:

1. **Callaway and Sant'Anna (2021)**: estimates separate an average treatment effect for each treatment cohort and each post-treatment time period:

$$\beta_g^\tau = \mathbb{E}[y_{i,t+\tau} - y_{i,t-1} | \text{Treatment cohort } g] - \mathbb{E}[y_{i,t+\tau} - y_{i,t-1} | \text{Not yet treated at time } t+\tau]$$

   In settings with multiple treatment cohorts, these cohort-specific average treatment effects can be averaged to give a single estimate, although in our placebo tests we will only have a single "treated" cohort. We use the doubly robust estimation method of Sant'Anna and Zhao (2020), and in some specifications we include time-invariant covariates. We also normalize the estimates for different horizons τ relative to $\tau = -1$.

2. **Imputation** (Borusyak et al., 2024): is a two-step procedure which starts by estimating a TWFE regression using only untreated observations:

$$y_{i,t} = \alpha_i + \gamma_t + \epsilon_{i,t} \text{ where } i, t \text{ are never-treated or not-yet-treated}$$

   The predicted values from this regression, $\hat{y}_{i,t}$, are then used to estimate the average treatment effect:

$$\beta^\tau = \frac{1}{N} \sum_i (y_{i,t+\tau} - \hat{y}_{i,t+\tau}) \text{ where } i \text{ is treated}$$

   We normalize the estimates β^τ relative to the latest pre-treatment period, $\tau = -1$.

3. **Wooldridge difference-in-differences** (Wooldridge, 2025): extends TWFE to include fully saturated interactions between the treatment-time indicator and group dummies:

$$y_{i,t} = \alpha_i + \gamma_t + \sum_{\tau=-T}^{T} \sum_g \beta_g^\tau \delta_{i,t-\tau} \mathbf{1}(\text{treatment cohort } g) + \epsilon_{i,t}$$

   With multiple treated cohorts, one could average the group-specific coefficient esti-

mates by event time τ to recover dynamic treatment effects as in an event study, although in our placebo tests we only have a single treated cohort. As with baseline TWFE, we normalize estimates relative to the latest pre-treatment period.

4. **de Chaisemartin and D'Haultfœuille (2020)**: allows for treatment switching by comparing newly treated units to continuously designated controls (and comparing continuously treated units to leavers, although in our placebo tests we will only have movements from untreated to treated):

$$\beta = \mathbb{E}[y_{i,t} - y_{i,t-1}|\text{Newly treated in } t] - \mathbb{E}[y_{i,t} - y_{i,t-1}|\text{Untreated in } t, t-1]$$

   We use the estimation method of de Chaisemartin and D'Haultfœuille (2024) and normalize estimates relative to the latest pre-treatment period.

5. **Local projections** (Dube et al., 2025): estimates difference-in-differences separately by event time τ:

$$y_{i,t+\tau} - y_{i,t} = \beta^{\tau}\delta_{i,t} + \theta X_{i,t} + \epsilon_{i,t}$$

   Adding time-invariant covariates to this specification accounts for differential trends across units. Since the outcome variable is differenced relative to the latest pre-treatment period, the coefficient estimates represent the effect normalized relative to this period.

3.3 Matching estimators

Matching estimators work by comparing treated units to a subset of control units that are sufficiently similar based on a set of observable covariates. This contrasts with TWFE-like estimators, which compare treated units to all control units available in the data. By subsetting control units, matching can theoretically improve the counterfactual estimated

for treated units.[3] We examine two different matching approaches:

1. **Propensity score matching** (Rosenbaum and Rubin, 1983): estimates the treatment probability of each unit as a function of covariates:

$$\Pr(\text{Treatment}_i|X) = \theta X_{i,t} + \epsilon_{i,t}$$

   Given estimated propensity scores for each unit, the researcher can compare treated units to only those control unit candidates with sufficiently high propensity scores, ensuring that the comparison is between similar units. For our purposes, we designate the units with the five highest propensity scores as our controls, and we drop the rest before estimating TWFE as in our baseline specification.

2. **Coarsened exact matching** (Iacus et al., 2012): discretizes continuous covariates by dividing into bins and then matches each treatment unit to only control units in the same bin. As a default, we use the Sturges method to calculate the number of bins to use in discretizing each covariate. After selecting control units within the same bin as the treated unit, we estimate TWFE comparing treatment to control as in our baseline specification.[4]

3.4 Synthetic-control-like estimators

We also examine the family of estimators building on the synthetic control method. These estimators are typically fitted on covariates and pre-treatment data in order to construct a close counterfactual for each treated unit, and tend to feature more complex statistical algorithms than the methods described above. In some of these methods, like the original

[3]The matching process selects which units to compare, not how, so practitioners could theoretically implement any difference-in-differences method on the subsetted data. In our approach, we intentionally estimate TWFE after matching to offer a comparison that highlights potential gains from the matching process itself.

[4]When we match on multiple covariates, we require control units to be in the same bin as the treated unit for all covariates in order to be included in the TWFE regression.

synthetic control, the algorithm determines heterogeneous weights for control units and then the counterfactual is calculated as a weighted mean, while in other methods the algorithm estimates the counterfactual directly (e.g. matrix completion). In contrast to TWFE-like methods (which assign equal weight to all control units) or matching estimators (which use covariates to determine control weights), synthetic-control-like methods typically use the pre-treatment *outcomes* in assigning weight to control units. We consider five estimators in this category:

1. **Synthetic control** (Abadie et al., 2010): estimates treatment effects as the difference between a treated unit's outcomes and a weighted mean of control units' outcomes, where the weights $\mathbf{w}$ are chosen to minimize pre-treatment and covariate differences:

$$\beta^{\tau} = y_{\text{Treated},t+\tau} - \sum_{i \in \{\text{Control}\}} \hat{w}_i y_{i,t+\tau} \qquad \hat{\mathbf{w}} \equiv \arg\min_{\mathbf{w}} \left\| \mathbf{X}_{\text{Treated}} - \sum_{i \in \{\text{Control}\}} w_i \mathbf{X_i} \right\|$$

   where $\mathbf{X}$ is a vector of pre-treatment outcomes and/or covariates, and the weights $\hat{\mathbf{w}}$ are constrained to be non-negative and sum to one. We normalize the estimates β^{τ} relative to $\tau = -1$.

2. **Elastic net synthetic control** (Doudchenko and Imbens, 2016): extends the canonical synthetic control method to allow for negative weights, weights that do not sum to one, and a permanent additive difference between the treated unit and controls. It relies on elastic net regularization to constrain the weights and performs cross validation to tune the hyperparameters for this regularization. We fit this method to data where we have normalized the outcome relative to the last pre-treatment period for each unit, which results in an estimate for $\tau = -1$ that is close to, but not exactly equal to, zero.

3. **Generalized synthetic control** (Xu, 2017): first estimates a factor model using control units only, obtaining factor estimates $\hat{F}_t$ for each time period t; then it estimates

the factor loading λ_i for each treated unit i using only the pre-treatment period; and finally estimates the treatment effect as the difference from the counterfactual implied by the factor estimates:

$$\beta_i^\tau = y_{i,t+\tau} - \lambda_i \hat{F}_t \text{ where } i \text{ is treated}$$

Cross validation is used to select the optimal number of factors. Similar to elastic net synthetic control, we normalize by differencing the outcomes for each unit relative to the latest pre-treatment period before estimating, but this results in some non-zero values for $\tau = -1$.

4. **Matrix completion** (Athey et al., 2021): treats potential outcomes under no treatment as a large matrix with missing values for the set of treated units in the treatment period and uses methods from computer science to "fill in" the missing counterfactuals. We choose the hyperparameters of the matrix completion algorithm using cross-validation. As with the previous two methods, we normalize via differencing the outcomes for each unit relative to the latest pre-treatment period, but this results in some non-zero values for $\tau = -1$.

5. **Synthetic difference-in-differences** (Arkhangelsky et al., 2021): expands on the synthetic control method by not only choosing weights w_i for control units, but also choosing weights ϕ_t for time periods, estimating the treatment effect using weighted differences-in-differences:

$$\beta^\tau = \sum_{i \in \text{Treated}} \frac{1}{N_{Tr}} \left(y_{i,t+\tau} - \sum_{t'=-T}^{-1} \phi_{t'} y_{i,t'} \right) - \sum_{j \in \text{Control}} w_j \left(y_{j,t+\tau} - \sum_{t'=-T}^{-1} \phi_{t'} y_{j,t'} \right)$$

The unit weights are chosen to minimize differences (up to a constant) between treated and control units during the pre-treatment period, while the time weights are chosen to minimize pre-treatment and post-treatment differences (up to a constant) among

control units. This method estimates the treatment effect for a single cohort of treated units but can be adapted to handle staggered adoption (Ben-Michael et al., 2022). To generate event study estimates, we follow the procedure detailed in Clarke et al. (2023). Furthermore, since this method estimates time period weights, we do not normalize estimates to the latest pre-treatment period.

3.5 Data and implementation

For each estimator, we estimate placebo tests using randomly drawn "events" from a monthly, state-level panel dataset spanning 1990–2024. We estimate each placebo event study separately using five different outcome variables that are relevant to local labor market and regional event studies:

1. Unemployment rate from the Current Population Survey (CPS)
2. Labor force participation rate from the CPS
3. Seasonally adjusted employment growth rate from the Local Area Unemployment Statistics (LAUS) program
4. Employment growth rate from the Quarterly Census of Employment and Wages (QCEW)
5. House price index growth rate from Federal Housing Finance Agency (FHFA)

In addition to looping over estimator, outcome, and event, we also conduct placebo tests both with and without time-invariant covariates.[5] When including time-invariant covariates, we use foreign-born share as a representative covariate because the choice and number of covariates has a relative minor impact on estimator performance on net. In section 5.3, we include alternative covariates and find little difference across them relative to using just foreign-born share. Examples of the specifications we test include the following: TWFE

[5]Not all estimators take time-invariant covariates, which we denote in Table 1 with "n.a." and do not include in our covariate specifications.

when the outcome is the LAUS employment growth rate with no covariates; propensity score matching when the outcome is the unemployment rate and matching on foreign-born share; and synthetic control when the outcome is labor force participation rate with no covariates.

We construct a full set of placebo tests by taking the Cartesian product between 13 estimators, two covariate options, five outcomes, 50 states and the District of Columbia, and 31 year-months selected by randomly sampling one month in each calendar year from 1992 to 2022.[6] We iteratively estimate event studies over the same set of randomly selected "events," isolating dynamic treatment effects at each time horizon ranging from 12 months before to 12 months after the event. For methods where it is possible, we normalize effects during the estimation process based on the latest pre-treatment period.[7] We aggregate point estimates across time horizons and placebo events to produce a unique distribution for each specification.

In the end, we produce over 134,000 placebo event study estimates across all specifications of interest. Since we randomize placebo events, the true "effect" is zero on average, and the spread of each distribution reflects the efficiency of the corresponding estimator. Beyond evaluating the relative performance of estimators, we conduct supplementary analyses to assess the impact of choices conditional on estimator, including normalization, seasonal adjustment, and covariate choice. We also investigate the significance of the research setting. In the next immediate section, we discuss our main findings comparing the relative performance of estimators across outcomes.

[6]We exclude the first and last 24 months of data from sampling to ensure that we can produce point estimates over the entirety of the 24-month window of each event study.

[7]This normalization practice standardizes a reference point across methods, which facilitates clear, concrete interpretations of point estimates. However, as mentioned in the methodology section, we do not obtain perfectly normalized estimates for elastic net synthetic control, generalized synthetic control, and matrix completion; additionally, we do not impose any normalization for synthetic difference-in-differences.

4 Relative performance of estimators

We focus on **mean squared error (MSE)** to evaluate the performance of each difference-in-differences estimator. For each estimator, we calculate the average MSE across placebo events, including all time horizons post-event-date. Ranking estimators by MSE for each outcome reveals which estimators perform best, in the sense of having the lowest variance in practice, for the specific outcome.[8]

Overall, we find that there is no single "superior" method. In Table 2, we report the MSE of placebo distributions by estimator and outcome. We display the MSE for TWFE in the first row, and in subsequent rows we provide the MSE difference relative to TWFE for each estimator, normalized as a percentage of the TWFE MSE to facilitate comparison.

TWFE-like estimators all perform identically to TWFE across the five economic outcomes of interest. Although these methods differ in how they estimate treatment effects with multiple cohorts of treated units, having only one treated unit in our setting simplifies them all to TWFE, which explains the identical performance. Matching estimators based on TWFE also perform similarly: up to about 10 percent better than TWFE when the outcome is LAUS employment or house price index and up to around 20 percent worse for the other outcomes.

Performance varies substantially for synthetic-control-like estimators across outcomes, outperforming TWFE in some cases but dramatically underperforming in other cases. These differences in MSE are nontrivial. Certain combinations of estimator and outcome produce variance that is 30–60 percent lower than TWFE, but in the worst cases other combinations yield variance that is multiples of that of TWFE. Even within the set of synthetic-control-like methods, performance is not strictly ordered across estimators. For example, matrix completion performs best for three of the five outcomes but worse than other methods for the two remaining outcomes. The one exception is generalized synthetic control, which

[8]Since MSE measures the difference from zero, it captures both bias and variance; however, as these are placebo estimates the bias term is zero by definition (since the true estimate is zero), therefore our measured MSE reflects only the variance term.

performs substantially worse compared to TWFE for every outcome.

The variation in performance of synthetic-control-like methods illustrates a bias-variance trade-off inherent to using these methods. Unlike TWFE or matching, synthetic-control-like methods use information about pre-treatment outcomes to construct the counterfactual for treated units, but this poses a risk of overfitting. In some settings, using pre-treatment outcomes helps deliver lower variance of placebo estimates, but in other settings these methods overfit on the pre-treatment period, resulting in higher variance of placebo estimates.

In sum, we do not find that there is a single method that performs best for every outcome, nor even that there is a general ordering of methods. These findings highlight that the research context matters greatly in determining the precision of event study estimates, and researchers would benefit from examining performance of placebo estimates in their setting to determine the "best" method to use.

5 Choices conditional on estimator

In this section, we examine the role of choices that practitioners must make beyond the estimator that can affect performance. We find meaningful improvements from normalizing estimates relative to the pre-treatment period or removing seasonal patterns in time series data. In contrast, we find much less of an impact on the variance of estimates from controlling for different sets of time-invariant covariates. These results highlight that other choices that researchers make in setting up their analysis matter at least as much as the choice of estimator.

5.1 Normalization

Our default specification for most estimators normalizes the coefficient estimates relative to the latest pre-treatment period. Consequently, the point estimate for the month preceding treatment is zero by construction, and one can interpret point estimates at other time hori-

Table 2: Mean squared error of estimates by method and outcome

	RU	LFPR	LAUS Emp. (SA)	QCEW Emp. (NSA)	HPI
Baseline mean squared error of estimates					
Two-way fixed effects	1.556	2.18	1.57	3.189	5.426
Percent difference relative to TWFE					
Callaway and Sant'Anna	0	0	0	0	0
Imputation	0	0	0	0	0
Wooldridge difference-in-differences	0	0	0	0	0
De Chaisemartin & D'Haultfoeuille	0	0	0	0	0
Local projections	0	0	0	0	0
Propensity score matching	12	19	-1	9	-9
Coarsened exact matching	14	19	-1	9	-7
Synthetic control	-16	-10	8	-4	-60
Elastic net synthetic control	248	13	78	-9	-65
Generalized synthetic control	95	164	10662	216	2278
Matrix completion	-29	-28	27	-35	52
Synthetic difference-in-differences	-20	-11	8	-8	25

Note: First row reports mean squared error (MSE) for TWFE placebo estimates. Subsequent rows report the MSE for each estimator normalized to be a percentage of the TWFE MSE. Cells are colored based on the MSE relative to TWFE, with blue representing improvement and red representing worse performance; more extreme values are darker. Matching estimators use foreign-born share as a covariate; all other methods exclude covariates. RU refers to the unemployment rate. LFPR refers to the labor force participation rate. LAUS Emp. (SA) refers to the Local Area Unemployment Statistics employment growth rate that is seasonally adjusted. QCEW Emp. (NSA) refers to the Quarterly Census of Employment and Wages employment growth rate that is not seasonally adjusted. HPI refers to the house price index growth rate. Source: Bureau of Labor Statistics; Census Bureau; Federal Housing Finance Agency.

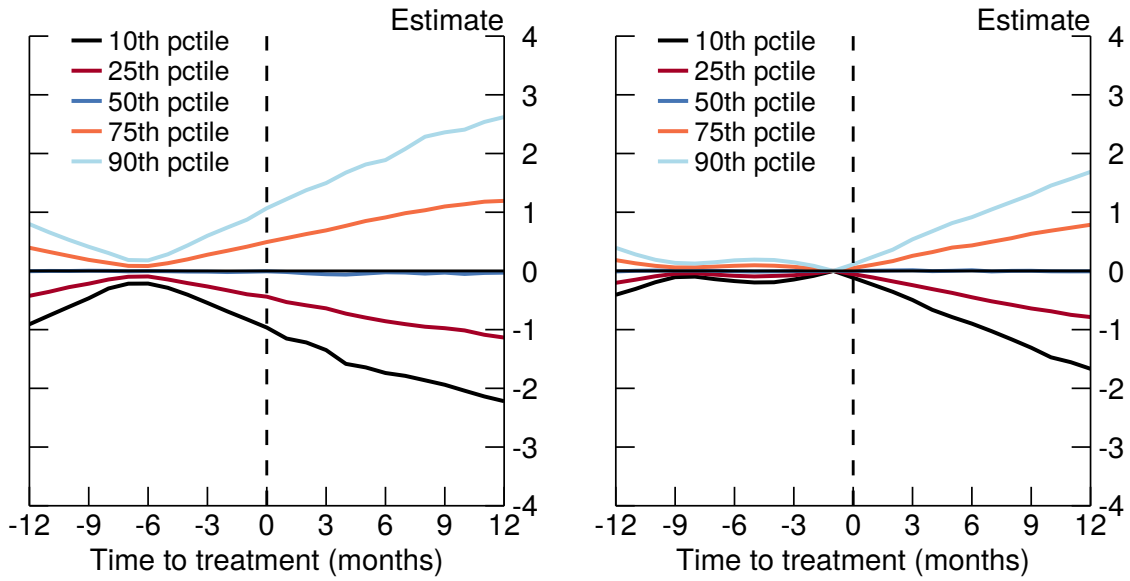
zons as measuring the change relative to this date. However, normalization is a choice that researchers can make (or not), and we examine how much this affects performance measured by the variance of placebo estimates.

We find that normalization vastly reduces the variance of estimates relative to estimating unnormalized specifications. Figure 1 shows a side-by-side comparison of placebo event study estimates without normalization (left) vs. with (right) for the specific example of synthetic control with the LAUS employment growth rate outcome. The distribution of placebo estimates without normalization is more diffuse, while the one with it is more concentrated around zero, reducing the MSE of the placebo estimates by 69 percent.[9] This example is representative of the gains from normalization in other combinations of estimators and outcomes, although the magnitude of gains may vary.

The benefits of normalization provide an explanation for how synthetic-control-like methods are able to outperform TWFE in some cases. Normalizing by the latest pre-treatment period is another way of using pre-treatment data to improve counterfactual estimation, as the synthetic-control-like methods do in various ways. Synthetic difference-in-differences notably estimates different weights for different pre-treatment periods, which nests our approach of normalizing by the latest pre-treatment period as a special case. However, while the performance of synthetic-control-like estimators varies relative to TWFE, we find improvement from normalizing in all settings in our analysis.

[9]We exclude the top and bottom deciles at each time horizon to prevent severe outliers from skewing MSE calculations.

Figure 1: Synthetic control estimates without normalization (left) vs. with (right)



Note: The outcome variable is the seasonally adjusted LAUS employment growth rate.
Source: Bureau of Labor Statistics; Census Bureau.

5.2 Seasonal adjustment

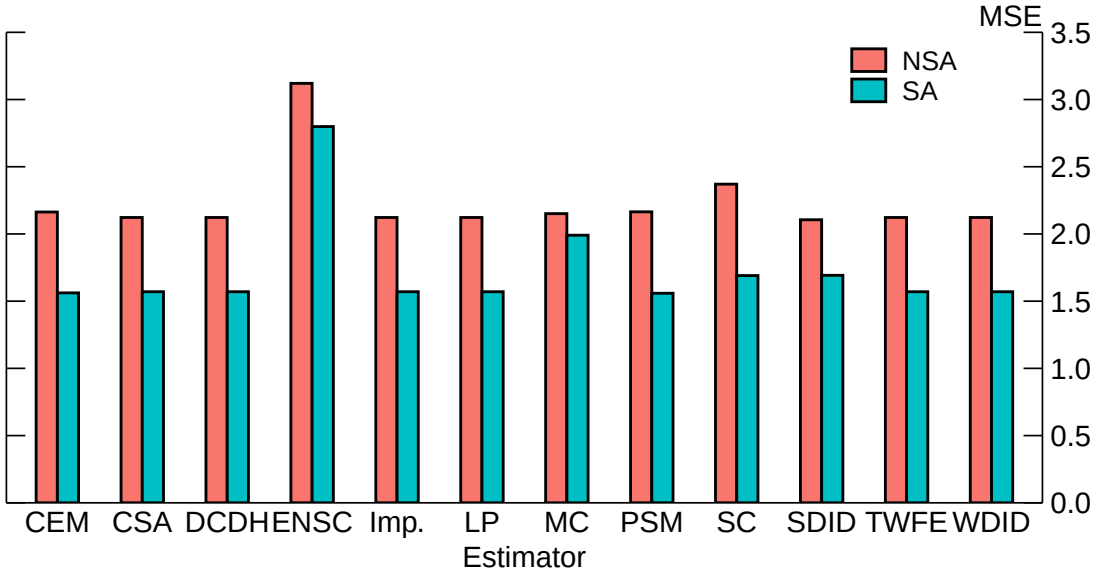
Using seasonal adjustment to filter out seasonal fluctuations in time series data offers another opportunity to improve performance. Figure 2 visualizes the MSE of estimates for each method by seasonal adjustment: the red bar indicates no seasonal adjustment, whereas the green bar indicates seasonal adjustment.[10] For each estimator, the relative heights of the different-color bars illustrate the extent to which seasonal adjustment lowers variance.

We find that seasonal adjustment noticeably lowers variance among placebo estimates. For the placebo estimates included in Figure 2, seasonal adjustment of LAUS employment data decreases MSE by an average of 23 percent across the estimators shown. The magnitude of reductions varies across these methods, but the difference between seasonally adjusted and not seasonally adjusted data is generally larger than the differences across estimators. The

[10]We show results only for LAUS employment growth rate because it is the only outcome variable with both seasonally adjusted and not seasonally adjusted data that are readily available at the state-level for our time range of interest. We omit placebo estimates using generalized synthetic control since this estimator is a severe outlier, and showing it would limit comparability.

systematic pattern of lower variance across estimators with seasonally adjusted data suggests that researchers should use seasonally adjusted data when possible.

Figure 2: Mean squared error of estimates by method and seasonal adjustment



Note: The outcome variable is the LAUS employment growth rate, and the predictor for matching estimators is foreign-born share. We exclude generalized synthetic control because of its sensitivity to outliers. From left to right, the estimators shown are coarsened exact matching (CEM), Callaway and Sant'Anna (CSA), D'Chaisemartin and D'Haultfœuille (DCDH), elastic net synthetic control (ENSC), imputation (Imp.), local projections (LP), matrix completion (MC), propensity score matching (PSM), synthetic control (SC), synthetic difference-in-differences (SDID), two-way fixed effects (TWFE), and Wooldridge difference-in-differences (WDID). The salmon bar denotes non-seasonally adjusted (NSA) data, whereas the teal bar denotes seasonal adjusted (SA) data.
Source: Bureau of Labor Statistics.

5.3 Covariates

Researchers may choose to control for a wide variety of covariates in difference-in-differences analyses; here we consider the choice of time-invariant controls.[11] Controlling for unit attributes, such as demographic composition or industry makeup, could absorb excess variance from confounders and thus increase comparability between states. Including covariates may improve precision and certain choices may do so better than others, so we investigate the im-

[11]Time-invariant controls account for persistent differences between treatment and control units. Some methods also allow for controlling for time-varying covariates, which we leave to future work to examine.

pact of covariate choice on MSE, conditional on estimator and outcome, for the six methods that allow for time-invariant covariates.

We use several different state-level characteristics taken from 1990 as the possible set of controls. These include average age, average income, Black population share, college-educated share, female employment rate, foreign-born share, manufacturing share of employment, and poverty rate, each calculated from the 1990 Census.

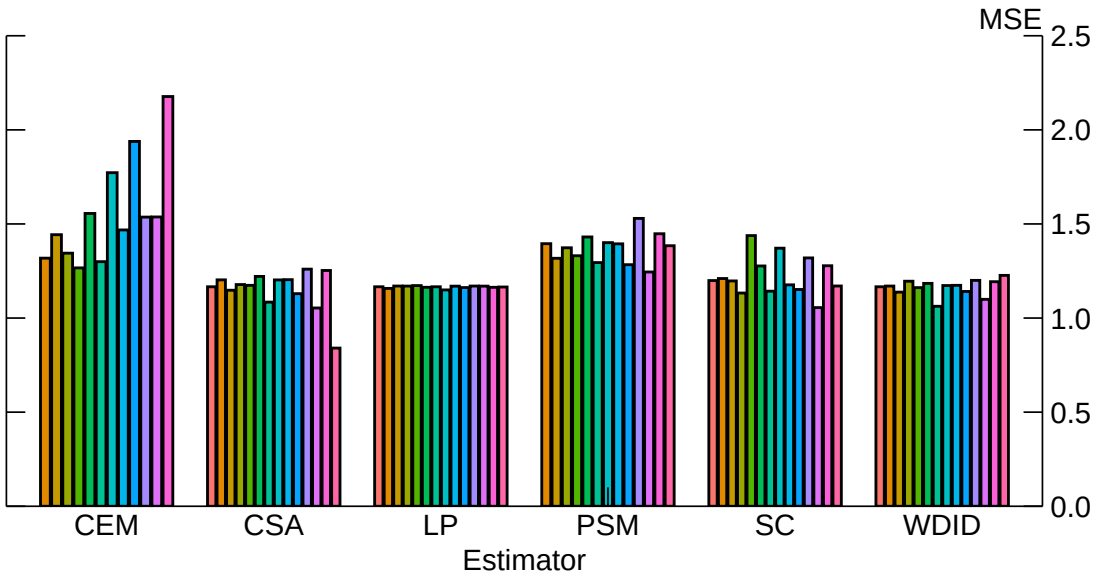
Figure 3 displays the MSE of estimates by method and covariate specification when the outcome is LAUS employment growth rate. From left to right within each set of bars, each colored bar corresponds to a unique set of covariates: the first bar includes no covariates, the next eight include one covariate each, then four with two randomly selected covariates, and finally one with four randomly selected covariates.[12]

We observe that the choice of at least one covariate generally has much less of an impact on precision than the choices for normalization or seasonal adjustment. Among the combinations of estimators and sets of at least one covariate, the vast majority result in MSEs within ± 10 percent of their respective estimator's baseline of no covariates. These minor differences in performance between choices of covariates are representative of other choices of the outcome variable.

One exception to this pattern is coarsened exact matching, where we see larger differences across different covariates. This method is the most sensitive to covariate choice because it strictly defines treatment and control comparisons based on the covariate cells, and sample size can shrink rapidly as more covariates are included in the matching process with the curse of dimensionality. Practitioners using this method should be aware of the sensitivity of this estimator to the choice of covariates.

[12]For the two matching estimators, we must includes covariates, so we omit the first specification for each.

Figure 3: Mean squared error of estimates by method and covariate specification



Note: The outcome variable is the seasonally adjusted LAUS employment growth rate. From left to right, the estimators shown are coarsened exact matching (CEM), Callaway and Sant'Anna (CSA), local projection (LP), propensity score matching (PSM), synthetic control (SC), and Wooldridge difference-in-differences (WDID). From left to right in each set of bars, each colored bars represents a unique covariate specification:

- None (except matching estimators)
- One covariate: average age; average income; Black population share; college-educated share; female employment rate; foreign-born share; manufacturing share of employment; poverty rate
- Two covariates: average income, average age; college-educated share, female employment rate; foreign-born share, Black population share; poverty rate, female employment rate
- Four covariates: average income, college-educated share, Black population share, manufacturing share of employment

Source: Bureau of Labor Statistics; Census Bureau.

6 Significance of the research setting

While researchers typically cannot choose the setting for a natural experiment, it may nonetheless matter as much as or more than the choices under researchers' control. States vary in their economic and demographic characteristics, and so the MSE of placebo estimates can vary substantially depending on which state is "treated." We show that MSE is systematically lower when the "treated" state is populous and diverse, and the variation across states is often much larger than variation across estimators. While differences in MSE across treated states align for most estimators, the optimal choice of control states varies substantially across estimators, since estimators form counterfactuals from control states differently. Our results highlight that which units are treated versus control makes a difference in precision at least as much as the choices within the researcher's control.

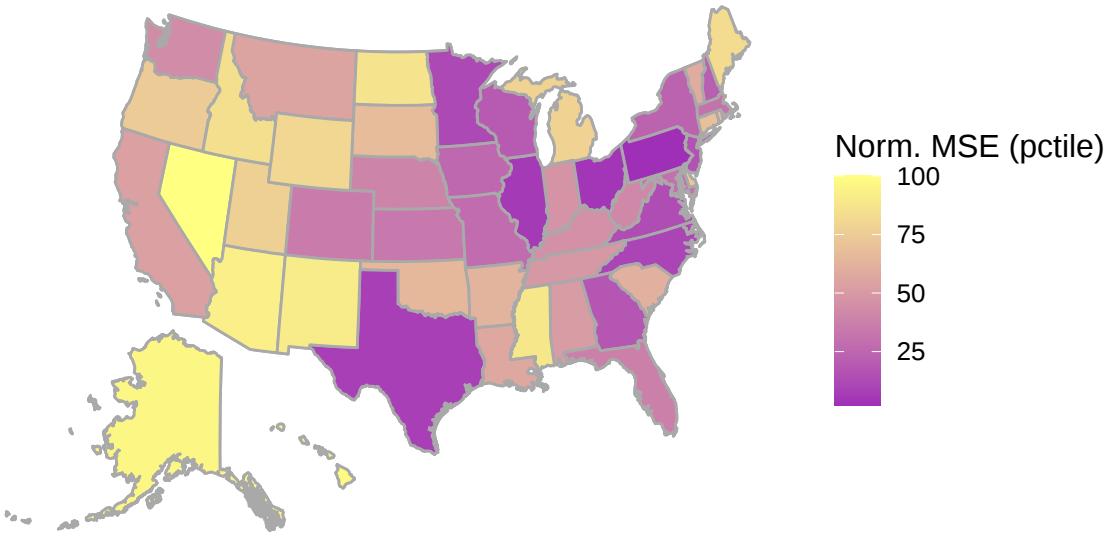
To facilitate comparison across states, giving equal weight to each estimator, we normalize MSE as follows. For each combination of estimator, outcome, and covariate choice, we average MSE by state across placebo events and convert these state-level averages to a unit scale so that the minimum value equals 0 and the maximum value equals 1.[13] We then average the normalized MSE for each state across specifications. In Figure 4, we visualize this measure, with color representing percentile of average normalized MSE. Purple indicates lower normalized MSE, while yellow indicates higher.

Overall, estimates are the most precise in populous, diverse states and least so in smaller and more distinctive ones. The five states with the highest precision are Pennsylvania, Ohio, Illinois, Texas, and North Carolina; the five states with the lowest precision are Nevada, Hawaii, Alaska, DC, and Arizona. These differences are substantial: the normalized MSE of Nevada is over 26 times that of Pennsylvania. In unreported results, we have confirmed that these patterns are common across estimators and are not skewed by any single estimator or specification in particular.

[13]The two covariate options are none or foreign-born share. Also, we exclude generalized synthetic control and elastic net synthetic control because they are extreme outliers.

The heterogeneity in performance across states highlights that practitioners should be mindful of where their event of interest occurred. Treated states that are populous and diverse resemble the contiguous United States more so than their idiosyncratic counterparts, thus yielding more precise estimates. If the event of interest is in a particularly distinctive state, such as Nevada, we encourage researchers to be mindful that estimates may vary more than in a typical setting.

Figure 4: Normalized MSE of estimates by state (percentile)



Note: We exclude generalized synthetic control and elastic net synthetic control because of their sensitivity to outliers. We obtain normalized MSE by calculating MSE by state within each specification (outcome × estimator × covariates), taking the percentage relative to the difference between the maximum and minimum values, and then averaging across specifications by state.
Source: Bureau of Labor Statistics; Census Bureau; Federal Housing Finance Agency.

In addition to heterogeneity across treated states, the choice of optimal control states may vary systematically as well. Four of our 13 estimators of interest construct counterfactuals by assigning weights to control states: propensity score matching, synthetic control, elastic net synthetic control, and synthetic difference-in-differences. Certain states may be better approximations of others and therefore would tend to receive more weight on average, so we examine which states each estimator tends to prefer designating as controls.

In Figure 5, we report the average control weight assigned to each state by method,

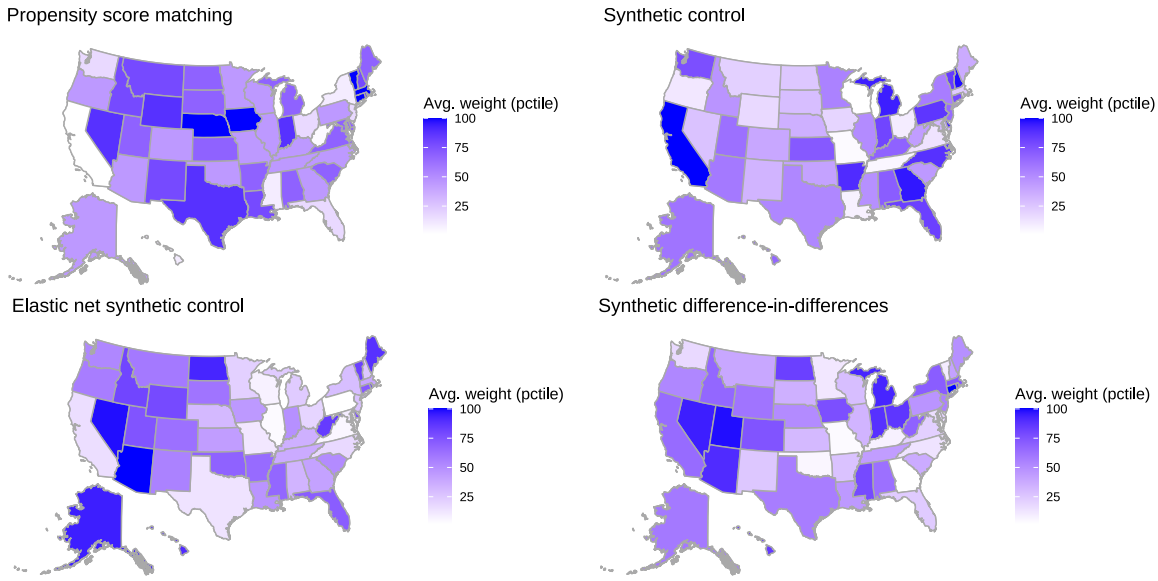
coloring by percentile based on averages across combinations of placebo events, covariates, and outcomes. Darker shades correspond to larger control weights, whereas lighter shades correspond to smaller ones. For any given state, one would interpret a larger average control weight as an indication of higher relative importance in constructing the counterfactual.

Unlike for the precision of estimates, we find that geographic patterns in control weights are not common across methods. For example, propensity score matching favors states in New England and the Great Plains, while synthetic difference-in-differences favors states in the Southwest and the Rust Belt.

Much of the difference in control weights likely stem from technical differences between these estimators, since they construct counterfactuals from control states differently. Propensity score matching weights based on propensity score for each state, which will tend to select states that are the most similar to the treated state on their own, while synthetic-control-like methods construct an aggregate counterfactual from multiple control states, even if none of the component states are particularly good matches for the treated state on their own. There is also substantial variation in control weight between each of the synthetic-control-like methods which reflects differences in how they assign weight, such as elastic net synthetic control allowing for negative weights and synthetic difference-in-differences weighting time periods in addition to control units.

The geographic variation in control weights conditional on estimator emphasizes that practitioners should be mindful of which states are represented in the data when considering which estimator to use. When only a subset of states is available, some estimators may be able to attain close to their optimal set of states, while other estimators' preferred states may not be in the subset. To construct the best possible counterfactual under data constraints, researchers should conduct placebo tests to identify and implement the estimator that performs the best given the particular set of control states represented in the data.

Figure 5: Average control weight by state (percentile)



Note: Weights are pooled across events, covariates, and outcomes.
Source: Bureau of Labor Statistics; Census Bureau; Federal Housing Finance Agency.

7 Conclusion

In this paper, we use placebo tests to evaluate the empirical performance of 13 difference-in-differences estimators across three categories: TWFE-like, matching, and synthetic-control-like. By producing event study estimates repeatedly over thousands of randomly drawn "events," we assess each estimator's efficiency based on the spread of placebo estimates.

We find that there is no single "superior" method for difference-in-differences. Synthetic-control-like methods sometimes outperform or underperform TWFE-like methods, depending on the estimator and outcome variable; the ranking of estimators by efficiency also varies across outcomes. Conditional on the estimator, normalization to the latest pre-treatment period and seasonal adjustment of the outcome variable meaningfully improve performance on average. However, which unit is "treated" matters just as much as which estimator is chosen: large, diverse states systematically yield lower variance in placebo estimates than their idiosyncratic counterparts.

Taken together, our placebo tests demonstrate that the performance of a difference-

in-differences method is highly context-dependent. The varied performance of synthetic-control-like methods across estimators and outcomes suggests that the practitioner must grapple with a bias-variance trade-off, in which optimizing on pre-treatment outcomes helps capture heterogeneous dynamics between units in certain cases but may lead to overfitting in others. The dichotomy in performance between large, diverse states and their idiosyncratic counterparts also emphasizes the significance of heterogeneous economic dynamics across units, which the analyst often cannot fully account for.

Researchers should be aware of the context dependence of difference-in-differences methods and strategize accordingly. We encourage analysts to normalize and seasonally adjust their outcome variables where possible. We also advise researchers to interpret results and communicate findings carefully, given that difference-in-differences estimates may vary significantly depending on which estimator is chosen, which outcome is selected, which state is treated, and which control states are represented in the data. Most importantly, we urge practitioners to conduct their own placebo tests to understand the practical performance of their methods in their unique setting.

8 Bibliography

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505. 10.1198/jasa.2009.ap08746.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–4118. 10.1257/aer.20190159.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* 116 (536): 1716–1730. 10.1080/01621459.2021.1891924.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (6): 2121–2168. 10.1257/aer.103.6.2121.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2025. "Difference-in-Differences Designs: A Practitioner's Guide." *Papers*, https://ideas.repec.org//p/arx/papers/2503.13323.html.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2022. "Synthetic Controls with Staggered Adoption." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84 (2): 351–381. 10.1111/rssb.12448.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?*." *The Quarterly Journal of Economics* 119 (1): 249–275. 10.1162/003355304772839588.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *The Review of Economic Studies* 91 (6): 3253–3285. 10.1093/restud/rdae007.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. 10.1016/j.jeconom.2020.12.001.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84 (4): 772–793, https://ideas.repec.org//a/aea/aecrev/v84y1994i4p772-93.html.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs*." *The Quarterly Journal of Economics* 134 (3): 1405–1454. 10.1093/qje/qjz014.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–2996. 10.1257/aer.20181169.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2024. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *The Review of Economics and Statistics* 1–45. 10.1162/rest_a_01414.

Chodorow-Reich, Gabriel, John Coglianese, and Loukas Karabarbounis. 2019. "The Macro Effects of Unemployment Benefit Extensions: a Measurement Error Approach*." *The Quarterly Journal of Economics* 134 (1): 227–279. 10.1093/qje/qjy018.

Clarke, Damian, Daniel Pailañir, Susan Athey, and Guido Imbens. 2023. "Synthetic Difference In Differences Estimation." February. 10.48550/arXiv.2301.11859, arXiv:2301.11859 [econ].

Doudchenko, Nikolay, and Guido W. Imbens. 2016. "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis."Technical Report w22791, National Bureau of Economic Research. 10.3386/w22791.

Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M. Taylor. 2025. "A Local Projections Approach to Difference-in-Differences." *Journal of Applied Econometrics* 1 (18): . 10.1002/jae.70000.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Hornbeck, Richard, and Enrico Moretti. 2024. "Estimating Who Benefits from Productivity Growth: Local and Distant Effects of City Productivity Growth on Wages, Rents, and Inequality." *The Review of Economics and Statistics* 106 (3): 587–607. 10.1162/rest_a_01208.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20 (1): 1–24. 10.1093/pan/mpr013.

Neumark, David, and William Wascher. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment." *American Economic Review* 90 (5): 1362–1396. 10.1257/aer.90.5.1362.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55. 10.2307/2335942.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics* 235 (2): 2218–2244. 10.1016/j.jeconom.2023.03.008.

Sant'Anna, Pedro HC, and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of econometrics* 219 (1): 101–122.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006, Themed Issue: Treatment Effect 1.

Wooldridge, Jeffrey M. 2025. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Empirical Economics*. 10.1007/s00181-025-02807-z.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76. 10.1017/pan.2016.2.