AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data

Licheng Liu Massachusetts Institute of Technology
Ye Wang University of North Carolina at Chapel Hill
Yiqing Xu Stanford University

Abstract: *This paper introduces a simple framework of counterfactual estimation for causal inference with time-series cross-sectional data, in which we estimate the average treatment effect on the treated by directly imputing counterfactual outcomes for treated observations. We discuss several novel estimators under this framework, including the fixed effects counterfactual estimator, interactive fixed effects counterfactual estimator and matrix completion estimator. They provide more reliable causal estimates than conventional two-way fixed effects models when treatment effects are heterogeneous or unobserved time-varying confounders exist. Moreover, we propose a new dynamic treatment effects plot, along with several diagnostic tests, to help researchers gauge the validity of the identifying assumptions. We illustrate these methods with two political economy examples and develop an open-source package,* `fect`*, in both* `R` *and* `Stata` *to facilitate implementation.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/ZVC9W5.

The linear two-way fixed effects (TWFE) model is one of the most commonly used statistical routines in the social sciences to establish causal relationships using observational time-series cross-sectional (TSCS) data, or long panel data. Such models are a popular choice because they can potentially control for a large set of unobserved unit- and time-invariant confounders. However, recent research points to several important drawbacks of fixed effects (FE) models (Blackwell and Glynn 2018; Imai and Kim 2019). First, the strict exogeneity assumption they rely on is often unrealistic – it not only requires the absence of time-varying confounders, but also rules out the possibility that past outcomes directly affect current treatment assignment (*no feedback*). It is well known that violations of strict exogeneity lead to biases in the causal estimates, yet methods for relaxing it or diagnosing its failure remain limited.

Second, TWFE models involve rigid functional form assumptions. When the treatments are dichotomous, TWFE models often assume their effects to be constant (*constant treatment effect*) and they affect the contemporaneous outcome only, not future outcomes (*no carryover effects*). Violation of the former will likely result in biased estimates even when strict exogeneity is satisfied, a problem receiving much attention in the literature recently. For example, de Chaisemartin and d'Haultfoeuille (2020) show that TWFE estimates are weighted averages of individualistic treatment effects, or treatment effects on each cell under the treatment condition. Because the weights can sometimes be negative due

Licheng Liu, PhD Candidate, Department of Political Science, Massachusetts Institute of Technology. 77 Massachusetts Avenue, MA 02139. Email: liulch@mit.edu. Ye Wang, Assistant Professor, Department of Political Science, University of North Carolina at Chapel Hill. 102 Emerson Dr, Chapel Hill, NC 27514. Email: yewang@unc.edu. Yiqing Xu, corresponding author. Assistant Professor, Department of Political Science, Stanford University. 616 Jane Stanford Way, Stanford, CA 94305. Email: yiqingxu@stanford.edu.

We thank Naoki Egami, Avi Feller, Neal Beck, Bernie Black, Dan de Kadt, Erin Hartman, Chad Hazlett, Danny Hidalgo, Apoorva Lal, Kosuke Imai, In Song Kim, Jeff Lewis, Marc Ratkovic, and seminar participants at NYU, UCSD, UCLA and MIT, as well as participants of PolMeth 2019 for helpful comments. We are grateful for the constructive comments by editors of AJPS and three anonymous reviewers, which we believe have helped us improve the paper significantly.

American Journal of Political Science, Vol. 00, No. 0, May 2022, Pp. 1–17

DOI: 10.1111/ajps.12723

to differential treatment timing and heterogeneous treatment effects, TWFE estimates may not even be convex combinations of the individualistic effects. In a staggered adoption setting where a treatment never switches back once it is on, Goodman-Bacon (2021) shows that negative weights are caused by temporal changes in the treatment effects of early treatment adaptors. Several papers aim to address this issue. For example, Strezhnev (2018) and Callaway and Sant'Anna (2021) suggest that under staggered adoption, researchers can instead estimate the average treatment effects for units that adopt the treatment at the same time, which they call the cohort average treatment effect; de Chaisemartin and d'Haultfoeuille (2020) propose to use observations only one period before or after the treatment's onset or exit, which leads to an estimator they call DID_M (M for 'multiple'). However, these approaches either have limited applicability by requiring a staggered adoption design or are statistically inefficient due to dropping many observations. Researchers so far have paid little attention to the no carryover effects assumption, though it is often testable by data.

In this paper, we introduce a simple framework that ameliorates these problems. We focus on TSCS data with dichotomous treatments, but they are allowed to switch back and forth (we call it a general panel treatment structure). Estimators under this framework take observations under the treatment condition as missing, use data under the control condition to build models and impute counterfactuals of treated observations based on the estimated models. We call them *counterfactual estimators*. This framework has several benefits. First, by not using the treated observations at the modelling stage and by imposing uniform weights on individualistic treatment effects on treated observations, it avoids the aforementioned negative weights problem and corrects biases induced by treatment effect heterogeneity. Second, it accommodates a variety of models, some of which can potentially relax the conventional strict exogeneity assumption. Third, it makes diagnostics and visualization much easier than with traditional TWFE models.

We discuss three methods under this framework, including (1) the fixed effects counterfactual (FEct) estimator, of which difference-in-differences (DID) is a special case; (2) the interactive fixed effects counterfactual (IFEct) estimator; and (3) the matrix completion (MC) estimator. Both IFEct and MC have recently emerged in the literature – see, for example, Gobillon and Magnac (2016) and Xu (2017) for the former and Kidziński and Hastie (2018) and Athey et al. (2021) for the latter. They are designed to construct a lower rank approximation of the outcome data matrix using information of untreated observations to account for potential time-varying confounders but differ in their ways of regularizing latent factors. Although FEct can be seen as a special case of IFEct, it is uniquely important because it provides a simple solution to the aforementioned weighting problem with the TWFE estimator. In addition to us, Borusyak, Jaravel, and Spiess (2021) and Gardner (2021) have independently proposed it as an improvement over TWFE. They call it the efficient estimator – because it is shown to be the most efficient among a class of linear unbiased estimators for the ATT – and the two-stage DID estimator, respectively.

Moreover, this paper aims to provide researchers with a set of diagnostic tools when making causal claims using TSCS data. A popular practice among researchers to evaluate the validity of the identifying assumptions is to draw a plot of the so-called 'dynamic treatment effects', which are coefficients of a series of interactions between a dummy variable indicating the treatment group – units that are exposed to the treatment for at least one period during the observed time window – and a set of time dummies indicating the time period relative to the onset of the treatment using a TWFE model. If these coefficients exhibit a monotonic trend leading toward the onset of the treatment, or a 'pretrend', the assumptions are deemed problematic. However, this method relies on parametric assumptions and the statistical tests derived from it are informal, often underpowered or even misleading (Roth 2020; Sun and Abraham 2021). Taking advantage of the counterfactual estimation framework, we improve the practice of estimating and plotting the dynamic treatment effects, or the average treatment effects on the treated (ATT) over different periods, without assuming treatment effect homogeneity of any kind.

In addition to visual inspections, we propose a set of statistical tests to help researchers evaluate the validity of the identifying assumptions. The core of these tests is based on a panel *placebo test*, in which we hide a few periods of observations right before the onset of the treatment for the treated units and use a model trained using the rest of the untreated observations to predict the untreated outcomes of those held-out periods. If the identifying assumptions are valid, the average differences between the observed and predicted outcomes in those periods should be close to zero. If, on the contrary, these differences are significantly different from zero, we obtain a piece of evidence that either the functional form assumption or strict exogeneity is likely invalid.

We then use this basic idea to construct two additional tests, a test for *no pretrend* and a test for *no carryover effects*. With the former, instead of hiding a few periods right before the treatment begins, we use

a leave-one-period-out approach to consecutively hide one pretreatment period (relative to the timing of the treatment) and repeatedly conduct a placebo test on observations in that period. By doing so, we have a more holistic view of whether the identifying assumptions will likely hold. The test for no carryover effects, on the other hand, is the mirror opposite of the placebo test in that it hides a few periods right after the treatment ends. If carryover effects do not exist, the average differences between the observed and predicted outcomes in those periods should be close to zero. This test is infeasible for the staggered adoption treatment structure, in which the treatment never switches back. However, under staggered adoption, potential carryover effects may not be concerns for researchers who care about the overall cumulative effects of the treatment over an extended period of time.

For all three tests, we use both a conventional difference-in-means (DIM) approach, which tests against the null of no difference, and an equivalence approach, which flips the null and tests against a prespecified difference. Consistent with the literature on equivalence tests in cross-sectional settings (Hartman and Hidalgo 2018; Hartman 2021), we show that the equivalence approach has advantages over the DIM approach when limited power is a concern. This is because as researchers collect more data, under valid identifying assumptions, it should be easier for them to declare equivalence by rejecting null in an equivalence test, not harder. Bilinski and Hatfield (2018), Dette and Schumann (2020) and Egami and Yamauchi (2021) propose similar tests recently in a DID setting. We recommend researchers consider the equivalence approach when data are limited.

This paper makes two main contributions to the literature. First, it introduces a counterfactual estimation framework to TSCS analysis that covers a variety of novel estimators. This new imputation approach addresses the weighting issue of TWFE models that causes concern for many researchers, and the new estimators introduced here can potentially control for decomposable time-varying confounders in a general panel data setting. Our second contribution is to develop a set of visualization and diagnostic tools to assist researchers in gauging the validity of the identifying assumptions and choosing the most suitable model for their applications.

This paper builds on earlier work on counterfactual estimation (or imputation methods) for causal inference. Heckman, Ichimura, and Todd (1997, 1998) first noted that, to identify the ATT, one only needs to impute counterfactuals for observations in the treatment group. This perspective has motivated a series of studies that try to predict the counterfactual in cross-sectional studies using various methods, such as regression (Lin 2013), the Oaxaca-Blinder estimator (Kline 2011) and machine learning algorithms (Künzel et al. 2019). The synthetic control method (SCM) first adopts the counterfactual approach in a panel setting (Abadie, Diamond, and Hainmueller 2010), but it is limited to comparative case studies, a specialized user case. We introduce it to systematically analysing panel/TSCS data.

We also contribute to an emerging literature on causal inference with panel/TSCS data and our approach has advantages over existing methods under various circumstances. Compared with existing factor-augmented methods (e.g. Gobillon and Magnac 2016; Xu 2017), which also use imputation methods, our framework can accommodate more complex TSCS designs, such as treatment reversal. Compared with TSCS methods based on matching and reweighting (e.g. Abadie 2005; Callaway and Sant'Anna 2021; de Chaisemartin and d'Haultfoeuille 2020; Imai and Kim 2019; Strezhnev 2018), our approach can accommodate more complex data structure and incorporate covariates more conveniently, and is often more efficient. It can also serve as a building block for doubly robust estimators, such as the augmented SCM (Ben-Michael, Feller, and Rothstein 2021).

This approach, of course, has limitations. First, the strict exogeneity assumption, which corresponds to baseline randomization, may be unrealistic in many applied settings, in which case researchers should consider methods based on sequential ignorability (Blackwell and Glynn 2018; Hazlett and Xu 2018; Imai, Kim, and Wang 2021). See Xu (2022) for a detailed discussion on the two identification regimes. With a general panel treatment structure, our method only allows limited carryover effects. Second, although we provide flexible modelling options, such as IFEct and MC, they are no panacea for all TSCS applications. The factor-augmented approach is more likely to suffer from biases due to model dependency and misspecification. Researchers have recently made efforts to alleviate this concern by proposing doubly robust estimators (e.g. Arkhangelsky et al. 2019; Ben-Michael, Feller, and Rothstein 2021); this paper does not incorporate these innovations because doing so would limit the applicability of our methods (e.g. by not allowing treatment reversal). Last but not least, the equivalence test approach requires users to specify an equivalence range, which may leave room for post hoc justification. Despite these drawbacks, we believe that the counterfactual imputation approach is a promising framework for TSCS analysis and can be extended to support a wide range of models.

Counterfactual Estimators

We first introduce the framework and the overall estimation strategy, and then discuss three novel estimators as examples.

A Simple Framework

Setup. Though our approach can accommodate both balanced and imbalanced panels, we consider a balanced panel with N units and T periods for notational convenience. Denote D_{it} the treatment status. Denote $Y_{it}(1)$ and $Y_{it}(0)$ the potential outcomes of unit i in period t when $D_{it} = 1$ and $D_{it} = 0$, respectively. Denote $\mathbf{X}_{it}$ a vector of the exogenous covariates, $\mathbf{U}_{it}$ the unobservable attributes and ε_{it} the idiosyncratic error term. Without loss of generality, we can define $\delta_{it} = Y_{it}(1) - Y_{it}(0)$ for unit i in period t. We assume the following class of outcome models for the untreated potential outcome:

Assumption 1 (Functional form). $Y_{it}(0) = f(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \varepsilon_{it}$, *in which* $f(\cdot)$ *and* $h(\cdot)$ *are known, parametric functions.*

Note that Assumption 1 requires additive separability of the four right-hand side terms. This class of models is scale dependent (Athey and Imbens 2006), e.g. transforming the outcome from levels to logarithms may render the identification assumptions discussed below invalid. It is easy to see that the classic two-group two-period DID approach assumes a model that is a special case in Assumption 1:

$$Y_{it}(0) = U_{it} + \varepsilon_{it} = \alpha_i + \xi_t + \varepsilon_{it}, \quad t = 1, 2,$$

in which α_i and ξ_t are unit and period FE. Hence, TWFE models' ability to control for unobserved confounders rests on the functional form assumption.

The setup, together with Assumption 1, rules out both temporal and spatial interference (Wang 2021), including potential anticipation effects and carryover effects. Borusyak, Jaravel, and Spiess (2021) show that the presence of anticipation effects will cause underidentification of the causal effects; the same logic applies to carryover effects. In a staggered adoption design, however, carryover effects are allowed because we can interpret δ_{it} as a combination of instant effect of the current treatment and cumulative carryover effects of past treatments on a treated unit relative to its potential outcome history under the never-treated condition – see Figure A3 on p. 6 in Supporting Information (SI) for a graphic illustration.

Estimands. The primary causal quantity of interest is the average treatment effect on the treated units, whose treatment status has changed at least once during the observed time window, that is,

$$ATT = \mathbb{E}[\delta_{it} | D_{it} = 1, C_i = 1]$$

in which $\delta_{it} = Y_{it}(1) - Y_{it}(0)$ by definition; and $C_i = 1$ if $\exists t, t'$ s.t. $D_{it} = 0$, $D_{it'} = 1$; otherwise, $C_i = 0$. For units that have never been exposed to the treatment condition, it is difficult to compute their treated potential outcomes without strong structural assumptions. Similarly, it is difficult to estimate causal effects on units that are always treated, and we drop them from the sample at the preprocessing stage.

In empirical work, researchers may be also interested in the average treatment effect on the treated at sth ($s > 0$) periods since the treatment's onset:

$$\begin{aligned} ATT_s = \mathbb{E}[\delta_{it} | & D_{i,t-s} = 0, \\ & \underbrace{D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1}_{s \text{ periods}}, \\ & C_i = 1], \quad s > 0. \end{aligned}$$

For the purpose of the diagnostic tests we will introduce later, we define $ATT_s = 0, \forall s \leq 0$. An alternative estimand used by de Chaisemartin and d'Haultfoeuille (2020) is the average instant treatment effect of changes in the treatment, that is,

$$\begin{aligned} AITC = \mathbb{E}[\delta_{it} | & (D_{i,t-1} = 0, D_{i,t} = 1) \quad \text{or} \\ & (D_{i,t} = 1, D_{i,t+1} = 0)]. \end{aligned}$$

It has the benefit of relaxing the no-carryover-effect assumption, but it is less applicable in many empirical applications because the effect of a treatment often takes time to manifest itself. For that reason, it is not the main focus of this paper. Our software package `fect` provides support for all the above estimands.

Assumption 2 (Strict exogeneity). $\varepsilon_{it} \perp\!\!\!\perp \{D_{js}, \mathbf{X}_{js}, \mathbf{U}_{js}\}$, *for all* i, $j \in \{1, 2, \ldots, N\}$ *and* $s, t \in \{1, 2, \ldots, T\}$.

Together with Assumption 1, Assumption 2 corresponds to baseline quasi-randomization conditional on $\mathbf{X}$ and $\mathbf{U}$, that is, $Y_{is}(0) \perp\!\!\!\perp D_{it} \,|\mathbf{X}_{i,1\ldots T}, \mathbf{U}_{i,1\ldots T}, \forall i, s, t$, in which $\mathbf{X}_{i,1\ldots T}$ and $\mathbf{U}_{i,1\ldots T}$ are the time series of $\mathbf{X}_{it}$ and $\mathbf{U}_{it}$, respectively. When $h(\mathbf{U}_{it}) = \alpha_i + \xi_t$ (as in DID), Assumption 2 implies the parallel trends assumption, that is,

$$\mathbb{E}[Y_{it}(0)|\mathbf{X}_{it}] - \mathbb{E}[Y_{is}(0)|\mathbf{X}_{is}] = \mathbb{E}[Y_{jt}(0)|\mathbf{X}_{jt}] - \mathbb{E}[Y_{js}(0)|\mathbf{X}_{js}], \quad \forall i, j, \forall t, s,$$

which states that, by expectation, the untreated potential outcome of all units follow parallel paths. When $\mathbf{U}_{it}$ is of a more general form, Assumption 2 implies

$$\mathbb{E}[Y_{it}(0)|\mathbf{X}_{it}, \mathbf{U}_{it}] - \mathbb{E}[Y_{is}(0)|\mathbf{X}_{is}, \mathbf{U}_{is}] = \mathbb{E}[Y_{jt}(0)|\mathbf{X}_{jt}, \mathbf{U}_{jt}] - \mathbb{E}[Y_{js}(0)|\mathbf{X}_{js}, \mathbf{U}_{js}], \quad \forall i, j, \forall t, s,$$

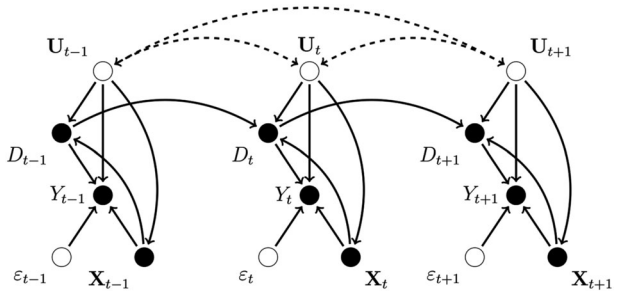
which states that conditional on the observed exogenous covariates and unobserved attributes (if we can extract them), the average changes in untreated potential outcome from period s to period t is the same between unit i and unit j. This leads to the third assumption.

Assumption 3 (Low-dimensional decomposition)**.** *There exists a low-dimensional decomposition of* $h(\mathbf{U}_{it}) : h(\mathbf{U}_{it}) = L_{it}$, *and* $rank(\mathbf{L}_{N\times T}) \ll \min\{N, T\}$. *For example,* $\mathbf{L} = \mathbf{\Lambda F}$, *in which* $\mathbf{\Lambda}$ *is an* $(N \times r)$ *matrix of factor loadings and* $\mathbf{F}$ *is an* $(r \times T)$ *matrix of factors and* $r \ll \min\{N, T\}$.

Assumption 3 allows us to condition on $\mathbf{U}_{it}$. To give a concrete example, if $\mathbf{U}_{it} = f_t \cdot \lambda_i$ is one dimensional, we can understand it as the impact of a common time trend f_t having a heterogeneous impact on each unit, whose heterogeneity is captured by λ_i. Moreover, when f_t is constant, $\mathbf{U}_{it}$ reduces to a set of unit FE; when λ_i is constant, it reduces to time FE. Hence, additive FE in DID models obviously satisfy this assumption. When unobserved confounders $\mathbf{U}_{it}$ exist, treatment assignment is dependent on observed untreated outcomes, thus, we are operating under a special case of missing not at random (Rubin 1976). Assumption 3 allows us to break this dependency by controlling for $\mathbf{U}_{it}$ approximated using data and can be understood as a feasibility assumption.

In Figure 1, we illustrate what the identifying assumptions entail using a directed acyclic graph (DAG). It shows that Assumptions 1 and 2 rule out anticipation effects or carryover effects (e.g. no arrows from D_t to Y_{t-1} or Y_{t+1}), feedback (e.g. no arrow from Y_{t-1} to D_t) and lagged dependent variables (no arrow from Y_{t-1} to Y_t); it also shows that the treatment effects of D_{it} on Y_{it} are separable from the influences of $\mathbf{U}_{it}$ and $\mathbf{X}_{it}$. This setup nests many existing models for TSCS data analyses, including TWFE and IFE models, although these models usually assume constant treatment effect, that is, $\delta_{it} = \delta$. If these assumptions are unsatisfied, research may turn to methods under sequential ignorability. See more discussion in Blackwell and Glynn (2018) and Imai and Kim (2019) on the potential tradeoffs.

FIGURE 1 A DAG Illustration



Notes: The figure presents a DAG (directed acyclic graph) consistent with Assumptions 1–3. Unit indices are dropped for simplicity. *Y*, *D*, *X*, ε represent the outcome, treatment, covariates and error term, respectively.

Estimation strategy. We define the observations under control and treatment conditions as $\mathcal{O} = \{(i, t)|D_{it} = 0\}$ and $\mathcal{M} = \{(i, t)|i \in \mathcal{T}, D_{it} = 1\}$, respectively, in which $\mathcal{O}$ stands for 'observed' and $\mathcal{M}$ stands for 'missing'. Although the outcome model researchers choose to employ may vary, estimation proceeds in a similar fashion with the following steps:

Step 1. On the subset of untreated observations ($\mathcal{O}$), fit a model of the response surface Y_{it}, obtaining $\hat{f}$ and $\hat{h}$. This step relies on the functional form assumptions on $f(\mathbf{X}_{it})$ and $h(\mathbf{U_{it}})$, as well as a lower rank representation of $\mathbf{U}$.
Step 2. Predict the counterfactual outcome $Y_{it}(0)$ for each treated observation using $\hat{f}$, $\hat{h}(\mathbf{U})$, that is, $\hat{Y}_{it}(0) = \hat{f}(\mathbf{X}_{it}) + \hat{h}(\mathbf{U}_{it})$, for all $(i, t) \in \mathcal{M}$.
Step 3. Estimate the individualistic treatment effects δ_{it} using $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for each treated observation $(i, t) \in \mathcal{M}$.
Step 4. Take averages of $\hat{\delta}_{it}$ to produce estimates for the quantities of interest. For example, for the ATT, $\widehat{ATT} = \frac{1}{|\mathcal{M}|}\sum_{\mathcal{M}} \hat{\delta}_{it}$; for the ATT at time period s since the treatment occurred $\widehat{ATT}_s = \frac{1}{|\mathcal{S}|}\sum_{(i,t)\in\mathcal{S}} \hat{\delta}_{it}$, in which $\mathcal{S} = \{(i, t)|D_{i,t-s} = 0, D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1\}$. $|\mathcal{A}|$ denotes the number of elements in set $\mathcal{A}$.

Because treated observations of early treatment adopters never serve as controls for late treatment adopters, we prevent the negative weights problem from its root cause (de Chaisemartin and d'Haultfoeuille 2020; Goodman-Bacon 2021). Compared with DID_M, our method is more efficient because it uses most available data without imposing stronger functional form assumptions.

Three Novel Estimators as Examples

In this subsection, we review three estimators as examples of this framework. They are conceptually similar because they follow the same identification strategy laid out above.

a) The FEct estimator. We start by introducing a counterfactual estimator in which $Y_{it}(0)$ is imputed based on a TWFE model, that is,

$$Y_{it}(0) = \mathbf{X}'_{it}\beta + \alpha_i + \xi_t + \varepsilon_{it}, \quad \text{for all } (i, t).$$

In other words, we assume $f(\mathbf{X}_{it}) = \mathbf{X}'_{it}\beta$ and $h(\mathbf{U}_{it}) = \alpha_i + \xi_t$. A linear constraint over the FE, $\sum_{D_{it}=0} \alpha_i = \sum_{D_{it}=0} \xi_t$, is imposed to achieve identification. This constraint also makes the grand mean parameter redundant.

It is easy to see that in a classic DID setup with two groups, two periods and no covariates, the FEct estimator is the DID estimator. How does FEct address the weighting issue with a general panel treatment structure? Arkhangelsky and Imbens (2021) show that with additive unit and time FE, any estimator that aims at identifying a convex combination of δ_{it} can be written as a weighted average of Y_{it}, where the weights $\{w_{it}\}_{1\le i\le N, 1\le t\le T}$ must satisfy the following four conditions: (1) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}D_{it} = 1$; (2) $\sum_{t=1}^{T} w_{it} = 0$ for any i; (3) $\sum_{t=1}^{N} w_{it} = 0$ for any t and (4) $w_{it}D_{it} \ge 0$ for any (i, t). Weights from both a TWFE model and FEct meet conditions (1)–(3). However, the former violates the last condition whereas latter does not; in fact, FEct imposes $w_{it:D_{it}=1} = \frac{1}{|\mathcal{M}|}$, which guarantees the identification of the causal quantities, such as ATT and ATT_s. We can therefore rewrite FEct as a weighting estimator; that is, each treated observation is matched with its predicted counterfactual $\hat{Y}_{it}(0) = \mathbf{W}^{(it)'}\mathbf{Y}_{\mathcal{O}}$, which is the weighted sum of all untreated observations. Comparison within each matched pair removes the biases caused by improper weighting that plague conventional FE models. We provide all the proofs in Section B in SI (pp. 12–15).

Proposition 1 (Unbiasedness and consistency of FEct)**.** *Under Assumptions 1–3, as well as some regularity conditions,*

$$\mathbb{E}[\widehat{ATT}_s] = ATT_s; \mathbb{E}[\widehat{ATT}] = ATT;$$

$$\widehat{ATT}_s - ATT_s \xrightarrow{p} 0; \quad \text{and} \quad \widehat{ATT} - ATT \xrightarrow{p} 0 \quad \text{as } N \to \infty.$$

Proposition 2 (FEct as a weighting estimator)**.** *Under Assumptions 1–3 and when there are no covariates, we have*

$$\widehat{ATT}_s = \frac{1}{|\mathcal{S}|}\sum_{(i,t)\in\mathcal{S}}[Y_{it} - \mathbf{W}^{(it)'}\mathbf{Y}_{\mathcal{O}}],$$

where $\mathbf{W}^{(it)'} = (\ldots, W_{js}^{(it)}, \ldots)_{(j,s)\in\mathcal{O}}$ *is a vector of weights that satisfy*

$$\sum_{(s:(i,s)\in\mathcal{O})} W_{is}^{(it)} = 1, \qquad \sum_{(j:(j,t)\in\mathcal{O})} W_{jt}^{(it)} = 1,$$

$$\sum_{(j:s\neq t,(j,s)\in\mathcal{O})} W_{js}^{(it)} = \sum_{(s:j\neq i,(j,s)\in\mathcal{O})} W_{js}^{(it)} = 0.$$

b) The IFEct estimator. FEct estimates will be biased when unobserved time-varying confounders exist. A couple of authors have proposed using factor-augmented models to relax the strict exogeneity assumption (Bai 2009; Bai and Ng 2021; Gobillon and Magnac 2016; Xu 2017). IFEct models the response surface of untreated potential outcomes using a factor-augmented model:

$$Y_{it}(0) = \mathbf{X}'_{it}\beta + \alpha_i + \xi_t + \lambda'_i f_t + \varepsilon_{it}, \quad \text{for all } (i, t).$$

In other words, $f(\mathbf{X}_{it}) = \mathbf{X}'_{it}\beta$ and $h(\mathbf{U}_{it}) = \alpha_i + \xi_t + \lambda'_i f_t$. When the model is correctly specified, IFEct is consistent.

Proposition 3 (Consistency of IFEct)**.** *Under Assumptions 1–3, as well as some regularity conditions,* $\widehat{ATT} \xrightarrow{p} ATT$ *as* $N, T \to \infty$.

c) The matrix completion estimator. Athey et al. (2021) introduce the MC method from the computer science literature as a generalization of factor-augmented models. Similar to FEct and IFEct, it treats a causal inference problem as a task of completing an $(N \times T)$ matrix with missing entries, where missing occurs when $D_{it} = 1$. Mathematically, MC assumes that the $(N \times T)$ matrix of $[h(\mathbf{U}_{it})]_{i=1,2,\ldots,N,t=1,2,\ldots,T}$ can be approximated by a lower rank matrix $\mathbf{L}_{(N\times T)}$, that is,

$$\mathbf{Y}(\mathbf{0}) = \mathbf{X}\beta + \mathbf{L} + \boldsymbol{\varepsilon},$$

in which $\mathbf{Y}$ is a $(N \times T)$ matrix of untreated outcomes; $\mathbf{X}$ is a $(N \times T \times k)$ array of covariates and $\boldsymbol{\varepsilon}$ represents an $(N \times T)$ matrix of idiosyncratic errors. As with IFEct, $\mathbf{L}$ can be expressed as the product of two r-dimension matrices: $\mathbf{L} = \boldsymbol{\Lambda}\mathbf{F}$. Unlike IFEct, however, the MC estimator does not explicitly estimate $\mathbf{F}$ and $\boldsymbol{\Lambda}$; instead, it seeks to directly estimate $\mathbf{L}$ by solving the following minimization problem:

$$\widehat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left[\sum_{(i,t)\in\mathcal{O}} \frac{(Y_{it} - L_{it})^2}{|\mathcal{O}|} + \lambda_L\|\mathbf{L}\|\right],$$

in which $\mathcal{O} = \{(i, t)|D_{it} = 0\}$, $\|\mathbf{L}\|$ is the chosen matrix norm of $\mathbf{L}$ and λ_L is a tuning parameter. Athey et al. (2021) propose an iterative algorithm to obtain $\widehat{\mathbf{L}}$ and show that $\widehat{\mathbf{L}}$ is an asymptotically unbiased estimator for

L. We summarize the algorithms for both IFEct and MC in SI (pp. 2–3).

Remark 1 (The difference between IFEct and MC)**.** *The main difference between IFEct and MC lies in the way they regularize the singular values when decomposing the residual matrix. IFEct uses a 'best subset' approach that selects the r biggest singular values, in which r is a fixed number and $r < \min\{N, T\}$, whereas MC imposes an L_1 penalty on all singular values with a tuning parameter λ_L (Figure 2). In the machine learning literature, they are referred to as hard impute and soft impute, respectively.*

Whether IFEct or MC performs better depends on context. In Section D.2 in SI (pp. 20–21), we provide Monte Carlo evidence to show that when the factors are strong and sparse, IFEct outperforms MC; otherwise, MC performs better. In practice, researchers may choose between the two models based on how they behave under the diagnostic tests we introduce in the next section. When $r = 0$ or when λ_L is bigger than the biggest singular value of the residual matrix, no factors are included in the model; as a result, IFEct or MC reduces to FEct.

The IFEct estimator was first proposed by Gobillon and Magnac (2016) in a DID setting where the treatment takes place at the same time for a subset of units. It is also closely related to the generalized SCM (Xu 2017), in which factors are estimated using the control group data only. In this paper, we accommodate with panel treatment structure, which allows treatment reversal. In other words, the generalized SCM can be seen as a special case of IFEct when the treatment does not switch back.

Remark 2 (Choosing the tuning parameters)**.** *In order to choose r for IFEct, we repeat Step 2 on a training set of untreated observations until $\hat{\beta}$ converges. The optimal r is then chosen based on a prespecified model performance metric, such as mean squared prediction error, using a k-fold cross-validation scheme. To preserve temporal correlations in the data, the test set consists of a number of triplets (three consecutive untreated observations of the same unit) from the treatment group. Similarly, for the MC estimator, we use k-fold cross-validation to select the λ_L. The test set is constructed in the same way as in IFEct.*

Remark 3 (Inferential methods)**.** *We rely on nonparametric block bootstrap and jackknife – both clustered at the unit level – to obtain uncertainty estimates for the treatment effect estimates. Our simulation results, reported in SI (pp. 16–17), suggest that both inferential methods work well with reasonable sample sizes (e.g. $T = 20, N = 50$). In practice, we recommend researchers use jackknife when the number of treated units is small.*

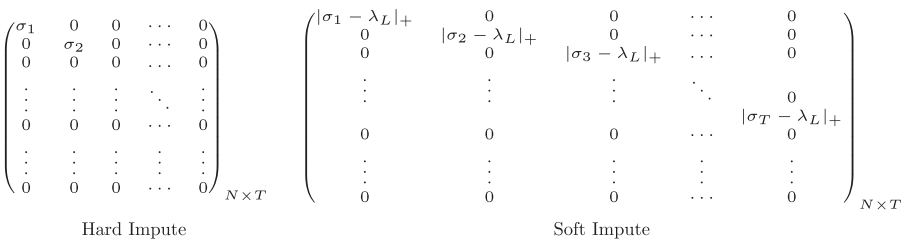
Diagnostics

In this section, we introduce a set of diagnostic tools to assist researchers probing the validity of the identifying assumptions. These assumptions should be considered collectively because strict exogeneity (Assumption 2) hinges on a correct functional form (Assumption 1) and bias removal is only possible when the feasibility condition (Assumption 3) is met. We first introduce a plot for dynamic treatment effects based on counterfactual estimators. We then propose several statistical tests for the implications of the identifying assumptions, including a placebo test, a test for no pretrend and a test for no carryover effects. The latter two can be seen as extensions of the placebo test.

A Plot for Dynamic Treatment Effects

In applied research with TSCS data, researchers often plot the so-called 'dynamic treatment effects', which are coefficients of the interaction terms between the treatment indicator and a set of dummy variables

FIGURE 2 Hard Impute (IFEct) vs. Soft Impute (MC)



Notes: The figure, adapted from Athey et al. (2021), illustrates how regularization works with IFEct (interactive fixed effects counterfactual) – which selects two factors in this case – and MC (matrix completion). It also shows that they are fundamentally similar ideas. $|a|_+ = \max(a, 0)$.

indicating numbers of periods relative to the onset of the treatment (lags and leads) – for example, $s = -4, -3, \ldots, 0, 1, \ldots, 5$ with $s = 1$ representing the first period a unit receives the treatment – while controlling for unit and time FE. Researchers then gauge the plausibility of the strict exogeneity assumption by eyeballing whether the coefficients in the pretreatment periods (when $s \leq 0$) exhibit an upward or a downward trend – often known as a 'pretrend' – or are statistically significant from zero. The magnitudes of the coefficients and corresponding p-values often depend on the baseline category researchers choose, which varies from case to case.

We improve the dynamic treatment effect plot by taking advantage of the counterfactual estimators. Instead of plotting the interaction terms, we plot the averages of the differences between Y_{it} and $\hat{Y}_{it}(0)$ for units in the treatment group ($C_i = 1$), re-indexed based on the time relative to the onset of the treatment. Specifically, we define $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$, for all $t, C_i = 1$. When the identifying assumptions are correct, it is easy to see that average pretreatment residuals will converge to zero, that is, $\widehat{ATT}_s \xrightarrow{p} 0$ for all $s \leq 0$.[1] Therefore, we should expect pretreatment residual averages to be bouncing around zero, that is, no strong pretrend. Figure 3 illustrates how we takes averages of $\hat{\delta}_{it}$ based on the timing relative to the next closest treatment.

This method has two main advantages over the traditional approach. First, it relaxes the constant treatment effect assumption. Even though the conventional dynamic treatment effect plot allows the treatment effects to be different across time, it assumes a constant effect for all treated units in a given time period (relative to the start of the next treatment).[2] Second, because a unit's untreated average has already been subtracted from $\hat{\delta}_{it}$, it is no longer necessary for researchers to choose a base category; to put it differently, the base category is set at a unit's untreated average after the time effects are partialed out. The dynamic treatment effects plot is an intuitive 'eyeball' test that can help researchers detect data and model issues instantly. However, it cannot differentiate the specific reasons why the identifying assumption may have failed, such as the anticipation effect, the presence of time-varying confounders or feedback from past outcomes.

[1] With some abuse of the terminology, we call the residual averages $\widehat{ATT}_s$ when $s \leq 0$.

[2] Sun and Abraham (2021) show that, under a staggered adoption design, if the dynamic treatment effects differ across cohorts, a spurious pretrend may arise even when the parallel trends assumption is valid.

We illustrate the plot using a simulated panel dataset of 200 units and 35 time periods based on the following data generating process (DGP) with two latent factors, f_{1t} and f_{2t}:

$$Y_{it} = \delta_{it} D_{it} + 5 + 1 \cdot X_{it,1} + 3 \cdot X_{it,2} + \lambda_{i1} \cdot f_{1t} + \lambda_{i2} \cdot f_{2t} + \alpha_i + \xi_t + \varepsilon_{it},$$

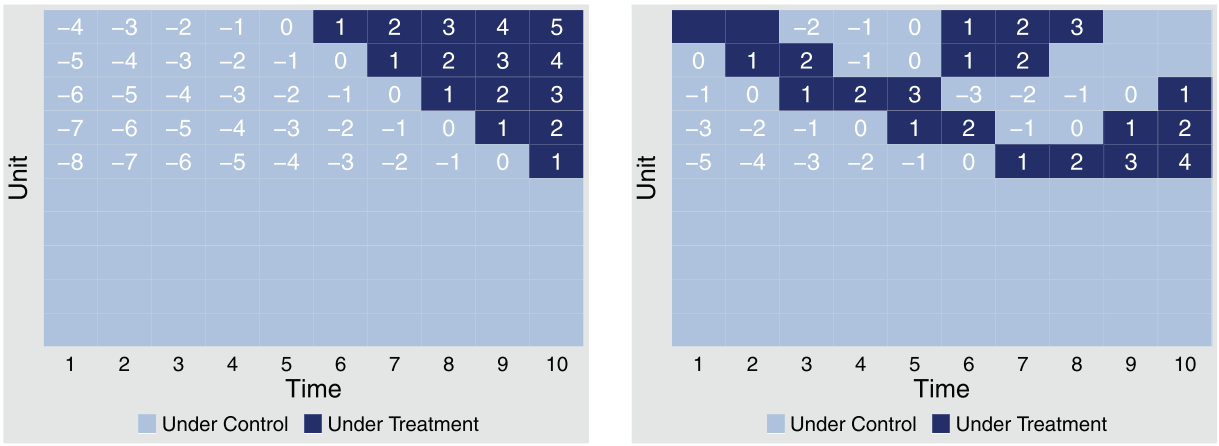
where the heterogeneous individualistic treatment effects are governed by $\delta_{it} = 0.4 s_t + e_{it}$ when $D_{it} = 1$, in which s_t represents the number of periods since the latest treatment's onset and e_{it} is i.i.d. $N(0, 0.16)$; and $\delta_{it} = 0$ when $D_{it} = 0$. This means the expected value of the treatment effect gradually increases as a unit takes up the treatment and there is no carryover effect. f_{1t} is a linear trend plus white noise and f_{2t} is an i.i.d. $N(0, 1)$ white noise. For each unit, the treatment may switch on and off. The probability of getting the treatment is dependent on the treatment status in the previous period as well as the interactive and additive FE (see p. 18 in SI for details; this DGP satisfies Assumptions 1–3). As a result, failure to adjust for these factors will lead to biases in the causal estimates.

Figure 4 shows the estimated dynamic treatment effects with 95% confidence intervals based on block bootstraps of 1000 times using the aforementioned counterfactual estimators. They are benchmarked against the true ATTs, which we depict with red dashed lines.

From the left panel of Figure 4, we see that using the FEct estimator, (1) a strong pretrend leads towards the onset of the treatment and multiple 'ATT' estimates (residual averages) in the pretreatment periods are significantly different from zero; and (2) there are sizeable positive biases in the ATT estimates in the post-treatment periods. We see a similar pattern in the post-treatment periods in the right panel where the MC estimator is applied, though with smaller biases. However, when using the IFEct estimator, the ATT estimates in both pretreatment and post-treatment periods are very close to the truth. This is expected because the DGP is generated by an IFE model with two latent factors and our cross-validation scheme picks the correct number of factors. To help researchers gauge the effective sample size, we plot the number of treated units at a given time period beneath the corresponding ATT estimate.

The dynamic treatment effects plot displays the temporal heterogeneity of treatment effects in an intuitive way. It is also a powerful visual tool for researchers to evaluate how plausible the identifying assumptions are. Next, we introduce several statistical procedures that formally test the implications of these assumptions. We start with a placebo test.

FIGURE 3 Estimating the Dynamic Treatment Effects



(a) Staggered Adoption (b) General Pattern

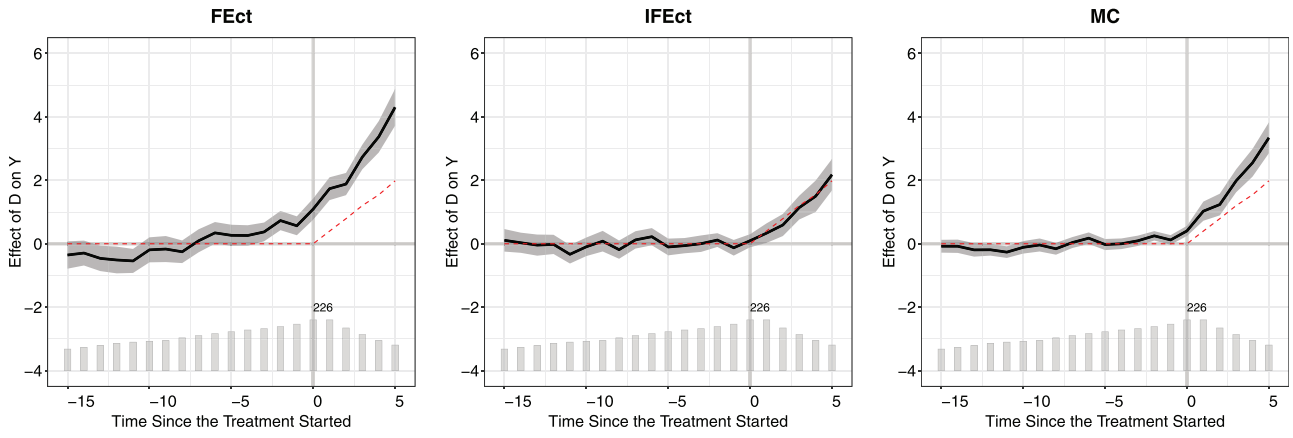
Notes: The figure shows the treatment status with two hypothetical examples: (a) staggered adoption and (b) a general panel treatment structure. Numbers correspond to time relative to the onset of a treatment. Several cells in (b) are not assigned numbers because left or right censorship of data makes their relative positions to a treatment uncertain.

A Placebo Test

The basic idea for the placebo test is straightforward: We assume that the treatment starts S periods earlier than its actual onset for each unit in the treatment group ($C_i = 1$) and apply the same counterfactual estimator to obtain estimates of ATT_s for $s = -(S-1), \ldots, -1, 0$. We can also estimate the overall ATT for the S pretreatment periods. If Assumptions 1–3 hold, we should expect the magnitude of this fake 'ATT' estimate is close to zero. If this 'ATT' estimate is statistically different from zero, we obtain a piece of evidence that some or all of the identifying assumptions are likely to be invalid.[3] For example, if a feedback effect from past outcome to current treatment exists (e.g. Y_{t-1} and D_t are positively correlated in

[3]In practice, S should not be set too large because the larger S is, the fewer pretreatment periods will remain for estimating the model. If both S and N_{tr} are too small, however, the test may be underpowered. In this and the following examples, we set $S = 3$.

FIGURE 4 Dynamic Treatment Effect for the Simulated Example



Notes: The above figure shows the dynamic treatment effects estimates from the simulated data using three different estimators: FEct, IFEct, and MC. The bar plot at the bottom of each panel illustrates the number of treated units at the given time period relative to the onset of the treatment (the number decreases as time goes by because there are fewer and fewer units that are treated for a sustained period of time). The red dashed lines indicate the true ATT.

Figure 1), which is a failure of the strict exogeneity assumption, it is likely to be detected by the placebo test given sufficient data.

Because a placebo test is a test for equivalence, as Hartman and Hildago (2018) point out, a simple DIM approach may suffer from limited power; that is, when the number of observations is small, failing to reject the null of the zero placebo effect does not mean equivalence holds. To address this concern, we introduce a variant of the equivalence test, where the null hypothesis is reversed:

$$ATT^p < -\theta_2 \quad \text{or } ATT^p > \theta_1,$$

in which $-\theta_2 < 0 < \theta_1$ are prespecified parameters, or equivalence thresholds. Rejection of the null hypothesis implies the opposite holds with a high probability, that is, $-\theta_2 \leq ATT^p \leq \theta_1$. In other words, if we collect sufficient data and show that the fake 'ATT' falls within a prespecified narrow range, we obtain a piece of evidence to support the validity of the identifying assumptions. $[-\theta_2, \theta_1]$ is therefore called the *equivalence range*. We use the two one-sided tests (TOST) to check the equivalence of ATT^p to zero. Following Hartman and Hildago (2018), we set $\theta_1 = \theta_2 = 0.36\hat{\sigma}_\varepsilon$, in which $\hat{\sigma}_\varepsilon$ is the standard deviation of the residualized untreated outcome[4]; alternatively, researchers may set the equivalence range based on an effect size they deem reasonable.

One advantage of the placebo test is that it is robust to model misspecification and immune from overfitting because it relies on out-of-sample predictions of $Y_{it}(0)$ in the placebo periods. Figure 5 shows the results from the placebo tests based on the three counterfactual estimators. We see that for FEct and MC, we can reject the null that the placebo effect is zero under the DIM test but cannot reject the null that the effect is outside the equivalence range – hence, equivalence does not hold – whereas IFEct behaves in the exact opposite way: The placebo effect is statistically indistinguishable from zero ($p = 0.534$), and we can reject the null hypothesis that the placebo effect is bigger than the true ATT ($p = 0.000$). Although the MC method fits the pretreatment periods well, it does not pass the placebo test using either the DIM approach ($p = 0.000$) or the equivalence approach ($p = 0.131$).

The main shortcoming of the equivalence approach is that researchers need to prespecify the equivalence range. $[0.36\hat{\sigma}_\varepsilon, 0.36\hat{\sigma}_\varepsilon]$ may be too lenient when the effect size is small relative to the variance of the residualized outcome. An alternative the literature suggests is to benchmark the minimum range against a reasonable guess of the effect size based on previous studies (e.g. Wiens 2001). However, such information is often unavailable. Because the ATT estimates from a TSCS analysis can be severely biased due to failures of the identification assumptions, unlike in experimental settings, they cannot provide valuable information for the true effect size, either. Moreover, setting the equivalence range in a post hoc fashion can lead to problematic results (Campbell and Gustafson 2018). The best practice would be for researchers to preregister a plausible effect size and use it to set the equivalence range before analysing data, as is a common practice in clinical trials.

Two Extensions

We now extend the placebo test to testing (1) whether a pretrend exists, especially when it takes place a few periods before the treatment starts and cannot be detected by the placebo test; and (2) whether the treatment has carryover effects.

A test for no pretrend. When a potential time-varying confounder is cyclical or does not present itself right before the treatment's onset, the placebo test may not be able to pick it up. Under this circumstance, we need a more global test for no pretrend. A natural approach is to jointly test a set of null hypotheses that the average of residuals for any pretreatment period is zero, that is, $ATT_s = 0$ for all $s \leq 0$ using an F test (see SI Section A.3, pp. 4–5, for details). However, because the test for no pretrend is also a test for equivalence, we develop an equivalence test with the following null:

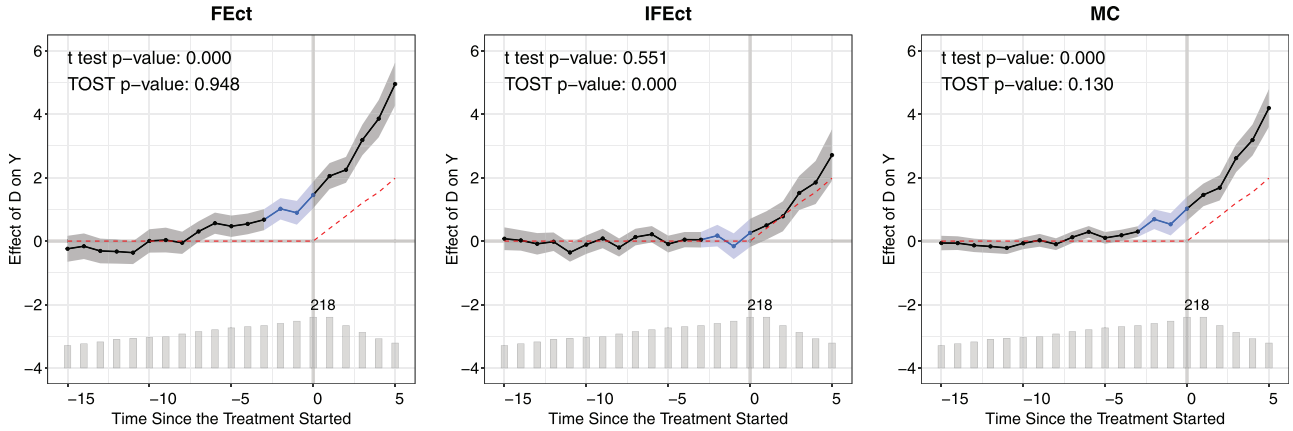
$$ATT_s < -\theta_2 \quad \text{or } ATT_s > \theta_1, \quad \forall s \leq 0,$$

in which $[-\theta_2, \theta_1]$ is the equivalence range. In other words, the null is considered rejected (hence, equivalence holds) only when the tests for all pretreatment periods generate significant results. This is clearly a conservative standard, as we are simultaneously testing multiple hypotheses; as a result, the Type I error will be smaller than the test size (e.g. 0.05).[5] The equivalence approach has an additional advantage over the F test in that when the sample size is large, a small confounder (or a few outliers) that only contributes to a neglectable amount of bias in the causal estimates will almost always cause rejection of

[4]Specifically, we run a TWFE model with time-varying covariates using untreated data only and calculate the standard deviation of the residuals. The literature maps it at a moderate effect size.

[5]Because the goal of an equivalence test is to control the Type I error, multiple testing, which makes the test more conservative, is not a major concern. See Hartman (2021) (footnote 11) for a discussion.

FIGURE 5 Placebo Tests for the Simulated Example



Notes: The figure shows the results of the placebo tests based on three different estimators: FEct, IFEct, and MC. The bar plot at the bottom of each panel illustrates the number of treated units at the given time period relative to the onset of the treatment. The red dashed lines indicate the true ATT. Three pretreatment periods ($s = -2, -1, 0$) serving as the placebo are rendered in blue. The p-values for the t test of the placebo effect and for the TOST are shown at the top-left corner of each panel. The equivalence range is set as $[-0.36\hat{\sigma}_\varepsilon, 0.36\hat{\sigma}_\varepsilon]$.

the null hypothesis of joint zero means using the F test. The equivalence test avoids this problem.

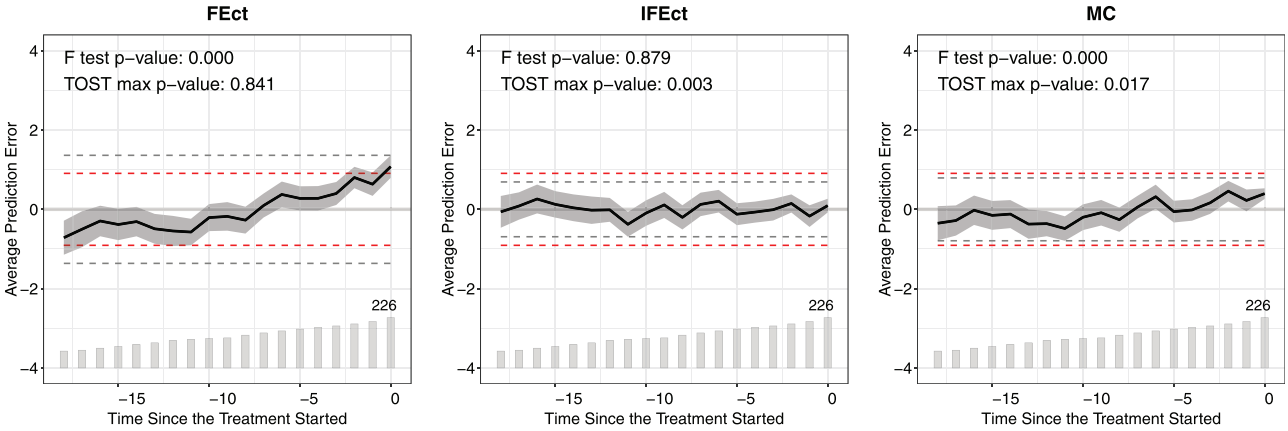
Building upon the basic idea of the placebo test, we use a leave-one-period-out approach to obtain an average out-of-sample prediction error for each period before the treatment's onset as long as data permits. Given a prespecified equivalence range, each of the TOST rejects the null of inequivalence when the bootstrapped one-sided confidence interval of pretreatment ATT_s (average prediction error in period s) falls within the range. In addition, we also calculate the *minimum range*, the smallest symmetric bound within which we can reject the null of inequivalence using our sample. In other words, the minimal range is determined by the largest absolute value of the range of the 90% confidence intervals of $\widehat{ATT}_{s,s\leq 0}$ in the pretreatment periods if we control the size $\alpha = 0.05$ (Hartman 2021). A rule of thumb is that when the minimum range is within the equivalence range, the test is considered passed. In Section D.3 in SI (pp. 21–22), we compare the performance of the F test and the equivalence test using simulations.

Figure 6 demonstrates the results of the equivalence test based on FEct, IFEct and MC using the simulated dataset. With FEct, the trend leading towards the onset of the treatment goes beyond the equivalence range and results in a wide minimum range. Therefore, we cannot reject the null that the pretreatment average prediction errors are beyond a narrow range – in other words, we cannot say that equivalence holds with high confidence. However, both IFEct and MC pass the test. The 90% confidence intervals of the pretreatment prediction error averages are within the equivalence range and the minimum range is narrower than the equivalence range. Note that the F test p-value for MC is .000, which points to potential model misspecification.

A test for no carryover effects. We extend the idea of the placebo test to testing the presence of carryover effects. Instead of hiding a few periods right before the treatment starts, we hide a few periods right after the treatment ends and predict $Y_{it}(0)$ in those periods. If carryover effects do not exist, we would expect the average prediction error in those periods to be close to zero. Once again, we use both the DIM approach and the equivalence approach. Figure 7 shows the results from applying this test to the simulated sample. Different from the dynamic treatment effects plot, the x-axis is now realigned based on the timing of the treatment's exit, not onset, for example, 1 represents one period after the treatment ends. The results show that the carryover effect does not seem to exist no matter which estimator or test is used, which is consistent with the DGP.

It is worth noting that the failure of the no carryover effects assumption does not necessarily invalidate our counterfactual estimation approach. If, by employing the proposed test, researchers find that the treatment effect persists after the treatment ends but in a limited time window, one strategy to proceed is to leave *a sufficient number* of periods after the end of the treatment as 'treated' and estimate the effects over these periods (as we do in the proposed test). Alternatively, researchers can change the definition of the treatment to '$D_{it} = 1$ if a unit has *ever* been under the treatment conditions, and $D_{it} = 0$ if otherwise', which essentially converts

FIGURE 6 Tests for No Pretrend: The Simulated Example



Notes: The above figure shows the results of the equivalence tests based on three different estimators: FEct, IFEct, and MC. Pretreatment average prediction errors and their 90% confidence intervals are drawn. The red dashed lines mark the equivalence range, whereas the grey dashed lines mark the minimum range. The bar plot at the bottom of each panel illustrates the number of treated units at the given time period relative to the onset of the treatment.

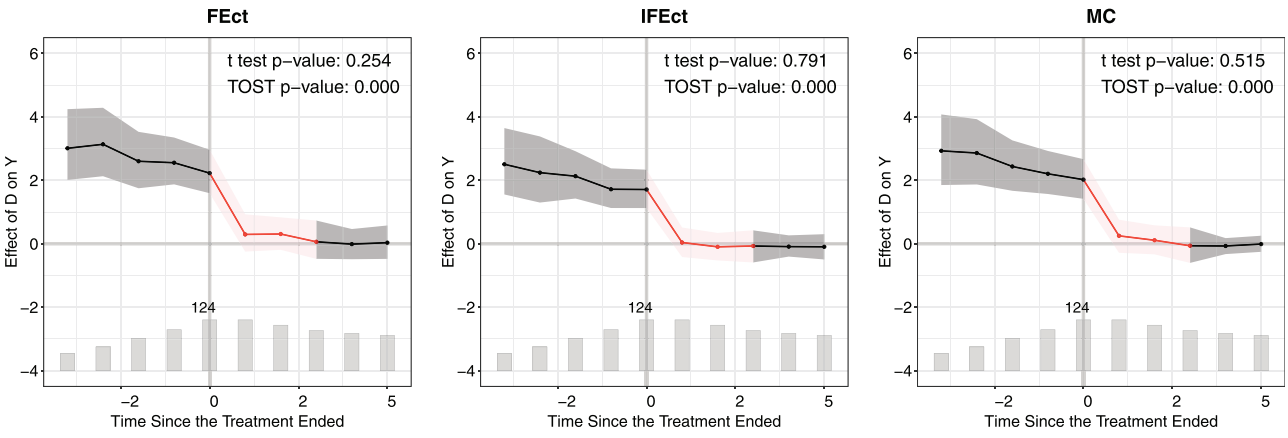
treatment assignment to a staggered adoption process, thus making the assumption for the no carryover effects unnecessary.

We summarize the diagnostic tests in Table 1. To lend support to the identifying assumptions, researchers can use either the DIM approach, if power is not a big concern, or the equivalence approach, if they have prior knowledge about the approximate effect size. No matter which approach researchers choose to use, visual inspection is always the first line of defense against erroneous causal claims based on invalid identifying assumptions.

Empirical Examples

We now apply the counterfactual estimators, as well as the diagnostics tools, to two empirical examples in political economy. The first example has a staggered adoption treatment structure whereas in the second one, the treatment switches back and forth. We start with FEct. If the results from FEct pass both the 'eyeball' test and the diagnostic tests, there is little need for more complex methods except for potential efficiency gains. If, however, the visual inspection or the tests suggest the identifying

FIGURE 7 Tests for No Carryover Effects Using the Simulated Example



Notes: The figure shows the results of the tests for no carryover effects based on three different estimators: FEct, IFEct, and MC. The bar plot at the bottom of each panel illustrates the number of treated units at the given time period relative to the end of the treatment. Three periods after the treatment ends are rendered in pink. The p-values for the *t* test of the carryover effects and for the TOST are shown at the top-right corner of each panel.

assumptions are unlikely to be true, we apply IFEct and MC and run diagnostics again. In both applications, we set $S = 3$ in the placebo tests. All uncertainty estimates are obtained using block bootstrap clustered at the unit level 1000 times.

Direct democracy and naturalization rates. Hainmueller and Hangartner (2019) study whether switching from direct democracy to indirect democracy increases naturalization rates for minority immigrants in Swiss municipalities using a staggered DID design. The outcome variable is minorities' naturalization rate in municipality i during year t. The treatment is a dummy variable indicating whether naturalization decisions are made by popular referendums. The dataset consists of 1211 Swiss municipalities over 19 years, from 1991 to 2009. The authors report that the naturalization rate increases by 1.339 percentage points on average (with a standard error of 0.161) after a municipality shifts the decision-making power from popular referendums to elected officials using a TWFE model.

We then apply FEct and obtain an estimate of 1.767 (with a standard error of 0.197), even larger than the original estimate. Plots for the dynamic treatment effects and placebo test are shown in Figure 8. We find that, first, the residual averages in the pretreatment periods are almost flat and around zero and the effect gradually takes off after the treatment begins. Second, with the placebo test, we cannot reject the null of zero placebo effect ($p = .422$), whereas we can reject the null whose magnitude is bigger than the default equivalence threshold ($p = .000$). Third, the F test does not reject the null of no pretrend at the 5% level ($p = .182$) while the TOST reject the null of inequivalence ($p = .001$). The test for carryover effects is not applicable because of the staggered adoption treatment structure. We also apply both IFEct and MC estimators to this example. It turns out that the cross-validation schemes find zero factors, in the case of IFEct, and a tuning parameter bigger than the first singular value of the residual matrix, in the case of MC, both of which imply maximum regularization (no factors). Hence, both methods reduce to FEct and give the exact same estimates as FEct.

In short, results from FEct are substantively the same as those from conventional TWFE models. However, counterfactual estimators like FEct allow us to check the validity of the identifying assumptions in a more convenient and transparent way.

Partisan alignment and grant allocation. Our second example is based on Fouirnaies and Mutlu-Eren (2015), in which the authors investigate whether partisan alignment between local councils in England and the central government bring localities more grants. The outcome of interest is the logarithm of specific grants per capita allocated to a local council. The treatment is a dummy variable indicating whether the government party controls the majority of local councils. The dataset spans 466 local councils from 1992 to 2012. The authors add locality-specific linear time-trends to a TWFE specification and find that partisan alignment increases specific grants allocated to a council – the increase peaks three years after alignment (see p. 24 in SI for the original figure). A TWFE model without the locality-specific trends, however, returns negative estimates for the effect of partisan alignment.

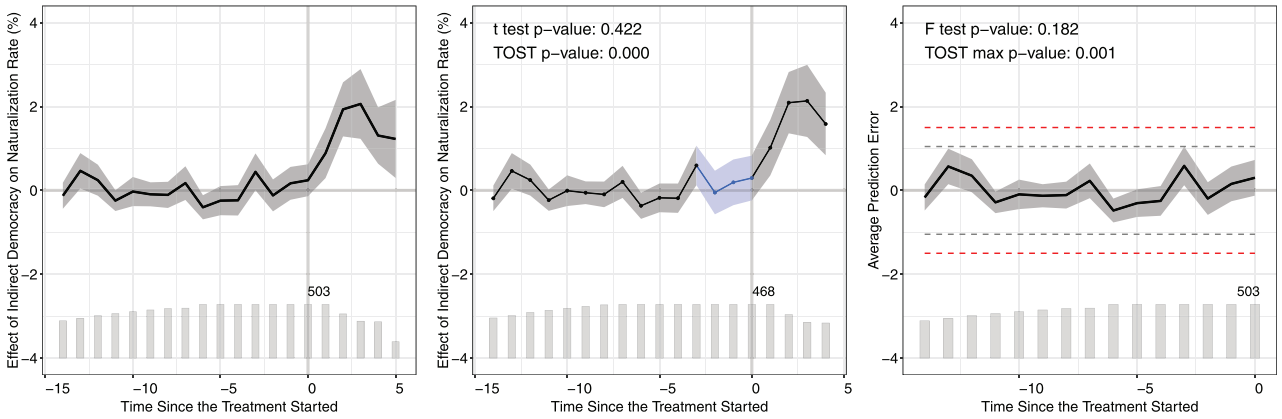
We apply the three estimators to the data and plot the estimated dynamic treatment effects in Figure 9a. It shows that, with FEct, the pretreatment residual averages consistently deviate from zero, suggesting potential violations of the identifying assumptions. With IFEct and MC, however, these averages are very close to zero. Figure 9b shows the results from the placebo test. With FEct, we cannot reject the null hypothesis of a non-zero

TABLE 1 Diagnostic Tests Summary

	Placebo Test		Testing (No) Pretrend		Testing (No) Carryover Effects	
	***t* test**	**TOST**	***F* test**	**TOST**	***t* test**	**TOST**
Null	$ATT^p = 0$	$\lvert ATT^p \rvert > \theta$	$ATT_s = 0, \forall s \leq 0$	$\lvert ATT_s \rvert > \theta, \exists s \leq 0$	$ACOE = 0$	$\lvert ACOE \rvert > \theta$
If rejecting the null	Invalidate assumptions	Support assumptions	Invalidate assumptions	Support assumptions	Invalidate no carryover	Support no carryover
Equivalence threshold θ		$0.36\hat{\sigma}_\varepsilon$ or eff		$0.36\hat{\sigma}_\varepsilon$ or eff		$0.36\hat{\sigma}_\varepsilon$ or eff

Notes: Both the t and F tests are conventional difference-in-means tests, testing against the null of no difference. 'Assumptions' refers to Assumptions 1–3 as a whole. $\hat{\sigma}_\varepsilon$ is the standard deviation of the residuals after two-way fixed effects are partialled out using untreated data only. ATT^p denotes the average placebo treatment effect on the treated. $ACOE$ denotes the average carryover effect. 'eff' represents an effect size that researchers deem reasonable.

FIGURE 8 The Effect of Indirect Democracy on Naturalization



(a) Dynamic Treatment Effects (b) Placebo Test (c) Testing No Pretrend

Notes: The figure shows the results from applying FEct to data from Hainmueller and Hangartner (2019). The left panel shows the estimated dynamic treatment effects using FEct. The middle panel shows the results from a placebo test using the 'treatment' in three pretreatment periods as a placebo. The right panel shows the results of an equivalence test for no pretrend, in which the red and grey dashed lines mark the equivalence range and the minimum range, respectively. The bar plot at the bottom of each panel illustrates the number of treated units at a given time period relative to the onset of the treatment.

placebo effect at the 5% level. With either IFEct or MC, both the DIM (t test) and the equivalence tests suggest the placebo effect is close to zero; however, IFEct seem to approximate the data better than MC as the pretrend looks almost completely flat and the estimated treatment effects are close to those in Figure 9a.[6] Finally, we report the results from the test for carryover effects in Figure 9c, in which we test the carryover effects up to five years after partisan alignment ends. Based on the result from IFEct, the test suggests that there are positive carryover effects at least three years after partisan alignment ends. As discussed earlier, violation of the no carryover effects assumption does not necessarily invalidate the research design, but suggests that a more flexible estimation strategy is required. After we remove observations in the three periods after the treatment ended in the model-building stage (Step 1 in the algorithm), we re-estimate the ATT and conduct the diagnostic tests again. The new results shown in Figure A15 in SI (p. 26) suggest that IFEct passes all diagnostic tests and is the most suitable model among the three. The magnitude of the effect remains similar.

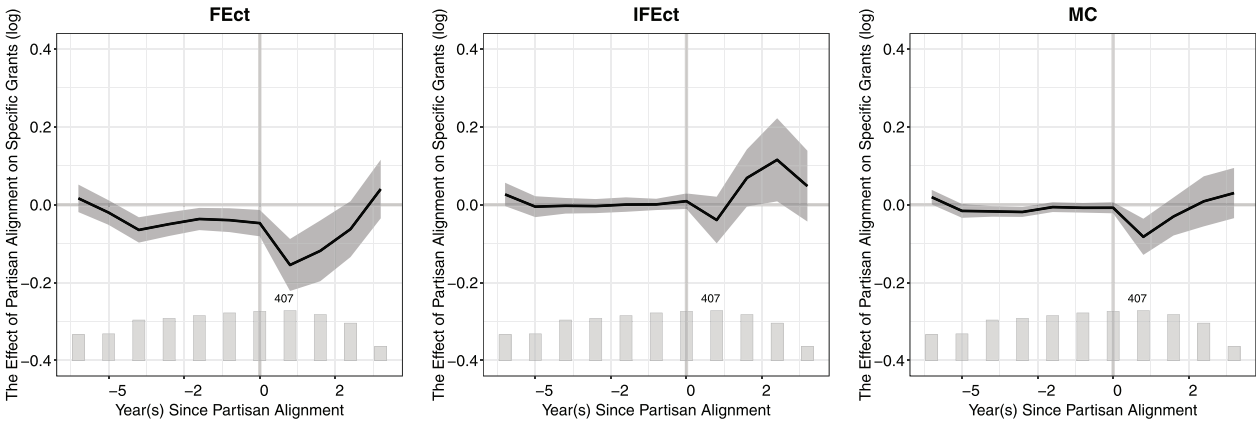
[6]The results from the equivalence tests are not greatly informative and are reported in Figure A14 in in SI (p. 25).

Conclusion

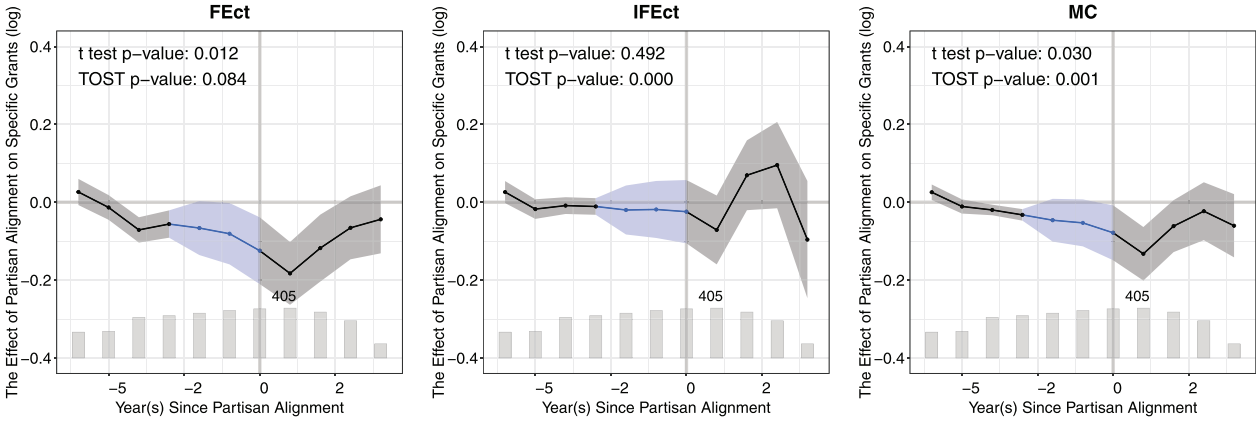
The commonly used TWFE models require strong assumptions to produce interpretable causal estimates; however, they remain highly valuable because of their versatility in accommodating different data structures and high computational efficiency. In this paper, we seek to improve current practices with TWFE models by providing a simple but powerful counterfactual estimation framework, the key to which can be described as 'fit data in the controls and impute counterfactuals to the treated', and by offering easy-to-implement diagnostic tests to assist researchers in probing the validity of the identifying assumptions.

We discuss three estimators under this framework, including FEct, IFEct and MC. It is important to note that IFEct and MC are not this paper's invention; they already exist in the literature. However, putting them in the same framework allows us to conduct diagnostics and evaluate their respective assumptions. Table 2 compares these estimators and other existing approaches and shows that they have several important advantages: They address the negative weights problem under heterogeneous treatment effects, accommodate general panel treatment structure without discarding data, can flexibly incorporate time-varying covariates and are amenable for diagnostic tests. In addition, IFEct and MC can account for decomposable time-varying confounders.

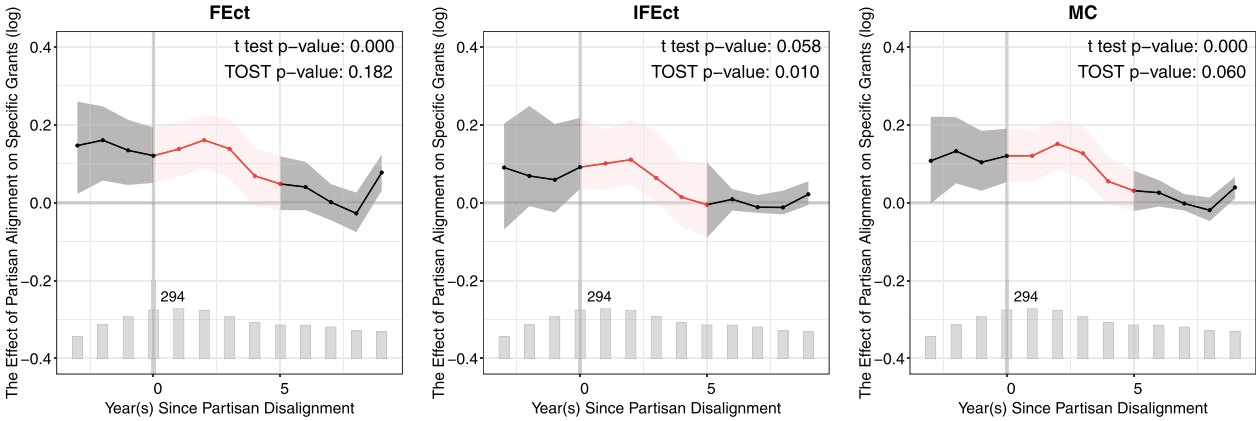
FIGURE 9 The Effect of Partisan Alignment on Grant Allocation



(a) Dynamic Treatment Effects



(b) Placebo Test



(c) Test for Carryover Effects

Notes: The blue dots in (b) represent the periods used in the placebo tests. The red dots in (c) represent the periods used in the tests for no carryover effects. The authors' original results based on a TWFE model with council-specific linear time trends are similar to the IFEct results.

TABLE 2 Comparison of Methods

	DID	wDID	DID_M	PM	TWFE	FEct	IFEct/MC
Accommodate heterogeneous treatment effects	x	x	x	x		x	x
Allow treatment reversal			x	x	x	x	x
Condition on time-invariant covariates		x		x			
Condition on time-varying covariates				x	x	x	x
Use most available data	x	x			x	x	x
Easy-to-implement diagnostic tests	x			x	x	x	x
Condition on $U_{it} = \lambda_i' f_t$							x

Notes: DID, wDID, DID_M, PM, TWFE, FEct, and IFEct/MC represent difference-in-differences, weighted difference-in-differences (Strezhnev 2018; Sun and Abraham 2021; Callaway and Sant'Anna 2021), multiple difference-in-differences (de Chaisemartin and d'Haultfoeuille 2020), panel match (Imai, Kim and Wang 2021), two-way fixed effects, fixed effects counterfactual, and interactive fixed effects counterfactural/matrix completion, respectively. $U_{it} = \lambda_i' f_t$ represents decomposable time-varying confounders.

We also improve the existing practice of estimating and plotting dynamic treatment effects and develop several statistical tests based on the new plot. These tests are based on out-of-sample predictions of untreated potential outcomes, and thus are immune to model misspecification or overfitting. We recommend researchers use the visual and statistical tests in a holistic manner to gauge the validity of the identifying assumptions, as we do with two empirical examples. Below we provide a checklist as a practical guide to analysing TSCS data using counterfactual estimators:

- Plot the treatment status of your data and ask whether strict exogeneity assumption is a plausible description of the treatment assignment process; if not, consider using methods based on sequential ignorability.
- Plot the outcome variable in a time-series fashion to spot outliers and irregularities; transform the data if necessary.
- Start with the simplest estimator, FEct, draw the dynamic treatment effects plot and perform both visual inspection and diagnostic tests (using either the DIM approach or the equivalence approach).
- If FEct does not pass the placebo test or the test for no pretrend, apply more complex models, such as IFEct and MC, and perform diagnostics again.
- If the chosen method fails the test for no carryover effects, remove several periods after the treatment ends from the model-building stage, then re-apply the method and conduct diagnostics again.
- Optionally, if a treatment effect is detected, perform subgroup analysis to understand which group(s) of units are driving the effect.
- Communicate your findings effectively, ideally with figures.

We provide two packages, `panelView` and `fect`, in both `R` and `Stata` to assist researchers in achieving these goals. We hope that this guide, as well as the tools we provide, will contribute to improved practices when researchers analyse TSCS data.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72(1): 1–19.

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.

Arkhangelsky, Dmitry, and Guido W. Imbens. 2021. "Double-Robust Identification for Causal Panel Data Models." NBER Working Paper, https://www.nber.org/papers/w28364.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2019. "Synthetic Difference in Differences." NBER Working Paper, https://www.nber.org/papers/w25532.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74(2):431–97.

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):1716–30.

Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77:1229–79.

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.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2021. "The Augmented Synthetic Control Method." *Journal of the American Statistical Association* 116(536):1789–803.

Bilinski, Alyssa, and Laura A. Hatfield. 2018. "Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions." Working Paper, https://arxiv.org/abs/1805.03273.

Blackwell, Matthew, and Adam N. Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* 112(4):1067–82.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." arXiv preprint arXiv:2108–12419.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225(2):200–30.

Campbell, Harlan, and Paul Gustafson. 2018. "What to Make of Non-Inferiority and Equivalence Testing with a Post-Specified Margin?" arXiv preprint arXiv:1807.03413.

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

Dette, Holger, and Martin Schumann. 2020. "Difference-in-Differences Estimation under Non-Parallel Trends." Working Paper, https://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/dette_schumann2020.pdf.

Egami, Naoki, and Soichiro Yamauchi. 2022. "Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Design." Political Analysis (forthcoming). Available at: https://doi.org/10.1017/pan.2022.8.

Fouirnaies, Alexander, and Hande Mutlu-Eren. 2015. "English Bacon: Copartisan Bias in Intergovernmental Grant Allocation in England." *Journal of Politics* 77(3):805–17.

Gardner, John. 2021. "Two-stage Differences in Differences." Boston College Working Paper, https://bit.ly/3AF32Af.

Gobillon, Laurent, and Thierry Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98(3):535–51.

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

Hainmueller, Jens, and Dominik Hangartner. 2019. "Does direct democracy hurt immigrant minorities? Evidence from naturalization decisions in Switzerland." *American Journal of Political Science* 63(3):530–47.

Hartman, Erin, and F. Daniel Hidalgo. 2018. "An equivalence approach to balance and placebo tests." *American Journal of Political Science* 62(4):1000–1013.

Hartman, Erin. 2021. "Equivalence Testing for Regression Discontinuity Designs." *Political Analysis* 29(4):505–21.

Hazlett, Chad, and Yiqing Xu. 2018. "Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data." Available at SSRN 3214231.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4):605–54.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2):261–94.

Imai, Kosuke, and In Song Kim. 2019. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data." *American Journal of Political Science* 63(2):467–90.

Imai, Kosuke, In Song Kim, and Erik H. Wang. 2021. "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science.* https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12685

Kidziński, Łukasz, and Trevor Hastie. 2018. "Longitudinal Data Analysis Using Matrix Completion." Working Paper, https://arxiv.org/abs/1809.08771.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101(3):532–37.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel and Bin Yu. 2019. "Meta learners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116(10):4156–65.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics* 7(1):295–318.

Roth, Jonathan. 2022. "Pre-Test with Caution: Event-Study Estimates After Testing for Parallel Trends." American Economic Review: Insights (forthcoming). Available at: https://www.aeaweb.org/articles?id=10.1257/aeri.20210236.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3):581–92.

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." Harvard University Working Paper, https://bit.ly/36kUJM6.

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

Wang, Ye. 2021. "Causal Inference under Temporal and Spatial Interference." Working Paper, https://arxiv.org/abs/2106.15074.

Wiens, Brian L. 2001. "Choosing an Equivalence Limit for Noninferiority or Equivalence Studies." *Controlled Clinical Trials* 23:2–14.

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

Xu, Yiqing. 2022. "Causal Inference with Time-Series Cross-Sectional Data: A Reflection." Working Paper, https://papers.ssrn.com/abstract=3979613.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Algorithms and Tests
Appendix B: Proofs
Appendix C: Inferential Methods
Appendix D: Additional Monte Carlo Evidence
Appendix E: Additional Information on the Empirical Examples