American Economic Review 2021, 111(12): 4088–4118
https://doi.org/10.1257/aer.20190159

Synthetic Difference-in-Differences[†]

By Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager*

We present a new estimator for causal effects with panel data that builds on insights behind the widely used difference-in-differences and synthetic control methods. Relative to these methods we find, both theoretically and empirically, that this "synthetic difference-in-differences" estimator has desirable robustness properties, and that it performs well in settings where the conventional estimators are commonly used in practice. We study the asymptotic behavior of the estimator when the systematic part of the outcome model includes latent unit factors interacted with latent time factors, and we present conditions for consistency and asymptotic normality. (*JEL* C23, H25, H71, I18, L66)

Researchers are often interested in evaluating the effects of policy changes using panel data, i.e., using repeated observations of units across time, in a setting where some units are exposed to the policy in some time periods but not others. These policy changes are frequently not random—neither across units of analysis, nor across time periods—and even unconfoundedness given observed covariates may not be credible (e.g., Imbens and Rubin 2015). In the absence of exogenous variation researchers have focused on statistical models that connect observed data to unobserved counterfactuals. Many approaches have been developed for this setting but, in practice, a handful of methods are dominant in empirical work. As documented by Currie, Kleven, and Zwiers (2020), difference-in-differences (DID) methods have been widely used in applied economics over the last three decades; see also Ashenfelter and Card (1985); Bertrand, Duflo, and Mullainathan (2004); and Angrist

*Arkhangelsky: CEMFI, Madrid (email: darkhangel@cemfi.es); Athey: Graduate School of Business, Stanford University, SIEPR, and NBER (email: athey@stanford.edu); Hirshberg: Department of Quantitative Theory and Methods, Emory University (email: davidahirshberg@emory.edu); Imbens: Graduate School of Business and Department of Economics, Stanford University, SIEPR, and NBER (email: imbens@stanford.edu); Wager: Graduate School of Business, and of Statistics (by courtesy), Stanford University (email: swager@stanford.edu). Thomas Lemieux was the coeditor for this article. We are grateful for helpful comments and feedback from referees, as well as from Alberto Abadie, Avi Feller, Paul Goldsmith-Pinkham, Liyang Sun, Erik Sverdrup, Yiqing Xu, Yinchu Zhu, and seminar participants at several venues. This research was generously supported by ONR grant N00014-17-1-2131 and the Sloan Foundation. The R package for implementing the methods developed here is available at https://github.com/synth-inference/synthdid. The associated vignette is at https://synth-inference.github.io/synthdid/.

[†]Go to https://doi.org/10.1257/aer.20190159 to visit the article page for additional materials and author disclosure statements.

and Pischke (2008). More recently, synthetic control (SC) methods, introduced in a series of seminal papers by Abadie and coauthors (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010; 2015; Abadie and L'Hour 2016), have emerged as an important alternative method for comparative case studies.

Currently these two strategies are often viewed as targeting different types of empirical applications. In general, DID methods are applied in cases where we have a substantial number of units that are exposed to the policy, and researchers are willing to make a "parallel trends" assumption that implies that we can adequately control for selection effects by accounting for additive unit-specific and time-specific fixed effects. In contrast, SC methods, introduced in a setting with only a single (or small number) of units exposed, seek to compensate for the lack of parallel trends by reweighting units to match their pre-exposure trends.

In this paper, we argue that although the empirical settings where DID and SC methods are typically used differ, the fundamental assumptions that justify both methods are closely related. We then propose a new method, synthetic difference in differences (SDID), that combines attractive features of both. Like SC, our method reweights and matches pre-exposure trends to weaken the reliance on parallel trend type assumptions. Like DID, our method is invariant to additive unit-level shifts, and allows for valid large-panel inference. Theoretically, we establish consistency and asymptotic normality of our estimator. Empirically, we find that our method is competitive with (or dominates) DID in applications where DID methods have been used in the past, and likewise is competitive with (or dominates) SC in applications where SC methods have been used in the past.

To introduce the basic ideas, consider a balanced panel with N units and T time periods, where the outcome for unit i in period t is denoted by Y_{it}, and exposure to the binary treatment is denoted by $W_{it} \in \{0, 1\}$. Suppose moreover that the first N_{co} (control) units are never exposed to the treatment, while the last $N_{tr} = N - N_{co}$ (treated) units are exposed after time T_{pre}.[1] Like with SC methods, we start by finding weights $\hat{\omega}^{sdid}$ that align pre-exposure trends in the outcome of unexposed units with those for the exposed units, e.g., $\sum_{i=1}^{N_{co}} \hat{\omega}_i^{sdid} Y_{it} \approx N_{tr}^{-1} \sum_{i=N_{co}+1}^{N} Y_{it}$ for all $t = 1, \ldots, T_{pre}$. We also look for time weights $\hat{\lambda}_t^{sdid}$ that balance pre-exposure time periods with postexposure ones (see Section I for details). Then we use these weights in a basic two-way fixed effects regression to estimate the average causal effect of exposure (denoted by τ):[2]

$$(1) \quad \left(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{sdid} \hat{\lambda}_t^{sdid} \right\}.$$

In comparison, DID estimates the effect of treatment exposure by solving the same two-way fixed effects regression problem without either time or unit weights:

$$(2) \quad \left(\hat{\tau}^{did}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\alpha,\beta,\mu,\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \right\}.$$

[1] Throughout the main part of our analysis, we focus on the block treatment assignment case where $W_{it} = \mathbf{1}(\{i > N_{co}, t > T_{pre}\})$. In the closely related staggered adoption case (Athey and Imbens 2021) where units adopt the treatment at different times, but remain exposed after they first adopt the treatment, one can modify the methods developed here. See the Appendix for details.

[2] This estimator also has an interpretation as a DID of weighted averages of observations. See equations (7) and (8) below.

The use of weights in the SDID estimator effectively makes the two-way fixed effect regression "local," in that it emphasizes (puts more weight on) units that on average are similar in terms of their past to the target (treated) units, and it emphasizes periods that are on average similar to the target (treated) periods.

This localization can bring two benefits relative to the standard DID estimator. Intuitively, using only similar units and similar periods makes the estimator more robust. For example, if one is interested in estimating the effect of anti-smoking legislation on California (Abadie, Diamond, and Hainmueller 2010), or the effect of German reunification on West Germany (Abadie, Diamond, and Hainmueller 2015), or the effect of the Mariel boatlift on Miami (Card 1990, Peri and Yasenov 2019), it is natural to emphasize states, countries or cities that are similar to California, West Germany, or Miami respectively relative to states, countries, or cities that are not. Perhaps less intuitively, the use of the weights can also improve the estimator's precision by implicitly removing systematic (predictable) parts of the outcome. However, the latter is not guaranteed: If there is little systematic heterogeneity in outcomes by either units or time periods, the unequal weighting of units and time periods may worsen the precision of the estimators relative to the DID estimator.

Unit weights are designed so that the average outcome for the treated units is approximately parallel to the weighted average for control units. Time weights are designed so that the average posttreatment outcome for each of the control units differs by a constant from the weighted average of the pretreatment outcomes for the same control units. Together, these weights make the DID strategy more plausible. This idea is not far from the current empirical practice. Raw data rarely exhibit parallel time trends for treated and control units, and researchers use different techniques, such as adjusting for covariates or selecting appropriate time periods to address this problem (e.g., Abadie 2005, Callaway and Sant'anna 2020). Graphical evidence that is used to support the parallel trends assumption is then based on the adjusted data. SDID makes this process automatic and applies a similar logic to weighting both units and time periods, all while retaining statistical guarantees. From this point of view, SDID addresses pretesting concerns recently expressed in Roth (2018).

In comparison with the SDID estimator, the SC estimator omits the unit fixed effect and the time weights from the regression function:

$$\left(\hat{\tau}^{sc}, \hat{\mu}, \hat{\beta}\right) = \underset{\mu,\beta,\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{sc} \right\}. \tag{3}$$

The argument for including time weights in the SDID estimator is the same as the argument for including the unit weights presented earlier: The time weight can both remove bias and improve precision by eliminating the role of time periods that are very different from the posttreatment periods. Similar to the argument for the use of weights, the argument for the inclusion of the unit fixed effects is twofold. First, by making the model more flexible, we strengthen its robustness properties. Second, as demonstrated in the application and simulations based on real data, these unit fixed effects often explain much of the variation in outcomes and can improve precision. Under some conditions, SC weighting can account for the unit fixed effects on its own. In particular, this happens when the weighted average of the outcomes for the control units in the pretreatment periods is exactly equal to the average of outcomes for the treated units during those pretreatment periods. In practice, this

equality holds only approximately, in which case including the unit fixed effects in the weighted regression will remove some of the remaining bias. The benefits of including unit fixed effects in the SC regression (3) can also be obtained by applying the SC method after centering the data by subtracting, from each unit's trajectory, its pretreatment mean. This estimator was previously suggested in Doudchenko and Imbens (2016) and Ferman and Pinto (2019). To separate out the benefits of allowing for fixed effects from those stemming from the use of time weights, we include in our application and simulations this synthetic control with intercept DIFP (Doudchenko-Imbens Ferman-Pinto) estimator.

I. An Application

To get a better understanding of how $\hat{\tau}^{did}$, $\hat{\tau}^{sc}$, and $\hat{\tau}^{sdid}$ compare to each other, we first revisit the California smoking cessation program example of Abadie, Diamond, and Hainmueller (2010). The goal of their analysis was to estimate the effect of increased cigarette taxes on smoking in California (based on the data from Orzechowski & Walker 2005). We consider observations for 39 states (including California) from 1970 through 2000. California passed Proposition 99 increasing cigarette taxes (i.e., is treated) from 1989 onwards. Thus, we have $T_{pre} = 19$ pretreatment periods, $T_{post} = T - T_{pre} = 12$ posttreatment periods, $N_{co} = 38$ unexposed states, and $N_{tr} = 1$ exposed state (California).

A. *Implementing SDID*

Before presenting results on the California smoking case, we discuss in detail how we choose the SC type weights $\hat{\omega}^{sdid}$ and $\hat{\lambda}^{sdid}$ used for our estimator as specified in (1). Recall that, at a high level, we want to choose the unit weights to roughly match pretreatment trends of unexposed units with those for the exposed ones, $\sum_{i=1}^{N_{co}} \hat{\omega}_i^{sdid} Y_{it} \approx N_{tr}^{-1} \sum_{i=N_{co}+1}^{N} Y_{it}$ for all $t = 1, \ldots, T_{pre}$, and similarly we want to choose the time weights to balance pre- and postexposure periods for unexposed units.

In the case of the unit weights $\hat{\omega}^{sdid}$, we implement this by solving the optimization problem

$$\left(\hat{\omega}_0, \hat{\omega}^{sdid}\right) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min}\, \ell_{unit}(\omega_0, \omega), \tag{4}$$

where

$$\ell_{unit}(\omega_0, \omega) = \sum_{t=1}^{T_{pre}} \left(\omega_0 + \sum_{i=1}^{N_{co}} \omega_i Y_{it} - \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it} \right)^2 + \zeta^2 T_{pre} \|\omega\|_2^2,$$

$$\Omega = \left\{ \omega \in \mathbb{R}_+^N : \sum_{i=1}^{N_{co}} \omega_i = 1,\ \omega_i = N_{tr}^{-1} \text{ for all } i = N_{co}+1, \ldots, N \right\},$$

where $\mathbb{R}_+$ denotes the positive real line. We set the regularization parameter ζ as

$$\zeta = \left(N_{tr} T_{post}\right)^{1/4} \hat{\sigma} \qquad \text{with} \qquad \hat{\sigma}^2 = \frac{1}{N_{co}\left(T_{pre} - 1\right)} \sum_{i=1}^{N_{co}} \sum_{t=1}^{T_{pre}-1} \left(\Delta_{it} - \bar{\Delta}\right)^2, \tag{5}$$

where

$$\Delta_{it} = Y_{i(t+1)} - Y_{it}, \qquad \text{and} \qquad \bar{\Delta} = \frac{1}{N_{co}(T_{pre}-1)} \sum_{i=1}^{N_{co}} \sum_{t=1}^{T_{pre}-1} \Delta_{it}.$$

That is, we choose the regularization parameter ζ to match the size of a typical one-period outcome change Δ_{it} for unexposed units in the pre-period, multiplied by a theoretically motivated scaling $(N_{tr} T_{post})^{1/4}$. The SDID weights $\hat{\omega}^{sdid}$ are closely related to the weights used in Abadie, Diamond, and Hainmueller (2010), with two minor differences. First, we allow for an intercept term ω_0, meaning that the weights $\hat{\omega}^{sdid}$ no longer need to make the unexposed pre-trends perfectly match the exposed ones; rather, it is sufficient that the weights make the trends parallel. The reason we can allow for this extra flexibility in the choice of weights is that our use of fixed effects α_i will absorb any constant differences between different units. Second, following Doudchenko and Imbens (2016), we add a regularization penalty to increase the dispersion, and ensure the uniqueness, of the weights. If we were to omit the intercept ω_0 and set $\zeta = 0$, then (4) would correspond exactly to a choice of weights discussed in Abadie, Diamond, and Hainmueller (2010) in the case where $N_{tr} = 1$.

We implement this for the time weights $\hat{\lambda}^{sdid}$ by solving[3]

$$\left(\hat{\lambda}_0, \hat{\lambda}^{sdid}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\, \ell_{time}(\lambda_0, \lambda), \tag{6}$$

where

$$\ell_{time}(\lambda_0, \lambda) = \sum_{i=1}^{N_{co}} \left(\lambda_0 + \sum_{t=1}^{T_{pre}} \lambda_t Y_{it} - \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} \right)^2,$$

$$\Lambda = \left\{ \lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{pre}} \lambda_t = 1,\ \lambda_t = T_{post}^{-1} \text{ for all } t = T_{pre}+1, \ldots, T \right\}.$$

The main difference between (4) and (6) is that we use regularization for the former but not the latter. This choice is motivated by our formal results, and reflects the fact we allow for correlated observations within time periods for the same unit, but not across units within a time period, beyond what is captured by the systematic component of outcomes as represented by a latent factor model.

We summarize our procedure as Algorithm 1.[4] In our application and simulations we also report the SC and DIFP estimators. Both of these use weights solving (4) without regularization. The SC estimator also omits the intercept ω_0.[5]

[3] The weights $\hat{\lambda}^{sdid}$ may not be uniquely defined, as ℓ_{time} can have multiple minima. In principle our results hold for any argmin of ℓ_{time}. These tend to be similar in the setting we consider, as they all converge to unique 'oracle weights' $\bar{\lambda}^{sdid}$ that are discussed in Section IIIB. In practice, to make the minimum defining our time weights unique, we add a very small regularization term $\zeta^2 N_{co} \| \lambda \|^2$ to ℓ_{time}, taking $\zeta = 10^{-6}\hat{\sigma}$ for $\hat{\sigma}$ as in (5).

[4] Some applications feature time-varying exogenous covariates $X_{it} \in \mathbb{R}^p$. We can incorporate adjustment for these covariates by applying SDID to the residuals $Y_{it}^{res} = Y_{it} - X_{it}\hat{\beta}$ of the regression of Y_{it} on X_{it}.

[5] Like the time weights $\hat{\lambda}^{sdid}$, the unit weights for the SC and DIFP estimators may not be uniquely defined. To ensure uniqueness in practice, we take $\zeta = 10^{-6}\hat{\sigma}$, not $\zeta = 0$, in ℓ_{unit}. In our simulations, SC and DIFP with this minimal form of regularization outperform more strongly regularized variants with ζ as in (5). We show this comparison in Table 6.

ALGORITHM 1—SDID

Data: $\mathbf{Y}, \mathbf{W}$
Result: Point estimate $\hat{\tau}^{sdid}$

1. Compute regularization parameter ζ using (5);
2. Compute unit weights $\hat{\omega}^{sdid}$ via (4);
3. Compute time weights $\hat{\lambda}^{sdid}$ via (6);
4. Compute the SDID estimator via the weighted DID regression

$$\left(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{sdid} \hat{\lambda}_t^{sdid} \right\};$$

TABLE 1

	SDID	SC	DID	MC	DIFP
Estimate	−15.6	−19.6	−27.3	−20.2	−11.1
Standard error	(8.4)	(9.9)	(17.7)	(11.5)	(9.5)

Notes: Estimates for average effect of increased cigarette taxes on California per capita cigarette sales over 12 posttreatment years, based on SDID, SC, DID, MC, DIFP, along with estimated standard errors. We use the "placebo method" standard error estimator discussed in Section IV.

Finally, we report results for the matrix completion (MC) estimator proposed by Athey et al. (2021), which is based on imputing the missing $Y_{it}(0)$ using a low rank factor model with nuclear norm regularization.

B. *The California Smoking Cessation Program*

The results from running this analysis are shown in Table 1. As argued in Abadie, Diamond, and Hainmueller (2010), the assumptions underlying the DID estimator are suspect here, and the −27.3 point estimate likely overstates the effect of the policy change on smoking. SC provides a reduced (and generally considered more credible) estimate of −19.6. The other methods, our proposed SDID, the DIFP and the MC estimator are all smaller than the DID estimator with the SDID and DIFP estimator substantially smaller than the SC estimator. At the very least, this difference in point estimates implies that the use of time weights and unit fixed effects in (1) materially affects conclusions, and, throughout this paper, we will argue that when $\hat{\tau}^{sc}$ and $\hat{\tau}^{sdid}$ differ, the latter is often more credible. Next, and perhaps surprisingly, we see that the standard errors obtained for SDID (and also for SC, DIFP, and MC) are smaller than those for DID, despite our method being more flexible. This is a result of the local fit of SDID (and SC) being improved by the weighting.

To facilitate direct comparisons, we observe that each of the three estimators can be rewritten as a weighted average difference in adjusted outcomes $\hat{\delta}_i$ for appropriate sample weights $\hat{\omega}_i$:

$$\hat{\tau} = \hat{\delta}_{tr} - \sum_{i=1}^{N_{co}} \hat{\omega}_i \hat{\delta}_i \qquad \text{where} \qquad \hat{\delta}_{tr} = \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} \hat{\delta}_i. \tag{7}$$

DID uses constant weights $\hat{\omega}_i^{did} = N_{co}^{-1}$, while the construction of SDID and SC weights is outlined in Section IA. For the adjusted outcomes $\hat{\delta}_i$, SC uses unweighted treatment period averages, DID uses unweighted differences between average treatment period and pretreatment outcomes, and SDID uses weighted differences of the same:

$$\hat{\delta}_i^{sc} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it}, \tag{8}$$

$$\hat{\delta}_i^{did} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} Y_{it},$$

$$\hat{\delta}_i^{sdid} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \sum_{t=1}^{T_{pre}} \hat{\lambda}_t^{sdid} Y_{it}.$$

The top panel of Figure 1 illustrates how each method operates. As is well-known (Ashenfelter and Card 1985), DID relies on the assumption that cigarette sales in different states would have evolved in a parallel way absent the intervention. Here, preintervention trends are obviously not parallel, so the DID estimate should be considered suspect. In contrast, SC reweights the unexposed states so that the weighted of outcomes for these states match California preintervention as close as possible, and then attributes any postintervention divergence of California from this weighted average to the intervention. What SDID does here is reweight the unexposed control units to make their time trend parallel (but not necessarily identical) to California preintervention, then apply a DID analysis to this reweighted panel. Moreover, because of the time weights, we only focus on a subset of the preintervention time periods when carrying out this last step. These time periods were selected so that the weighted average of historical outcomes predicts average treatment period outcomes for control units, up to a constant. It is useful to contrast the data-driven SDID approach to selecting the time weights to both DID, where all pretreatment periods are given equal weight, and to event studies where typically the last pretreatment period is used as a comparison and so implicitly gets all the weight (e.g., Borusyak and Jaravel 2016; Freyaldenhoven, Hansen, and Shapiro 2019).

The lower panel of Figure 1 plots $\hat{\delta}_{tr} - \hat{\delta}_i$ for each method and for each unexposed state, where the size of each point corresponds to its weight $\hat{\omega}_i$; observations with zero weight are denoted by an $\times$ symbol. As discussed in Abadie, Diamond, and Hainmueller (2010), the SC weights $\hat{\omega}^{sc}$ are sparse. The SDID weights $\hat{\omega}^{sdid}$ are also sparse—but less so. This is due to regularization and the use of the intercept ω_0, which allows greater flexibility in solving (4), enabling more balanced weighting. Observe that both DID and SC have some very high influence states, that is, states with large absolute values of $\hat{\omega}_i(\hat{\delta}_{tr} - \hat{\delta}_i)$ (e.g., in both cases, New Hampshire). In contrast, SDID does not give any state particularly high influence, suggesting that after weighting, we have achieved the desired "parallel trends" as illustrated in the top panel of Figure 1 without inducing excessive variance in the estimator by using concentrated weights.

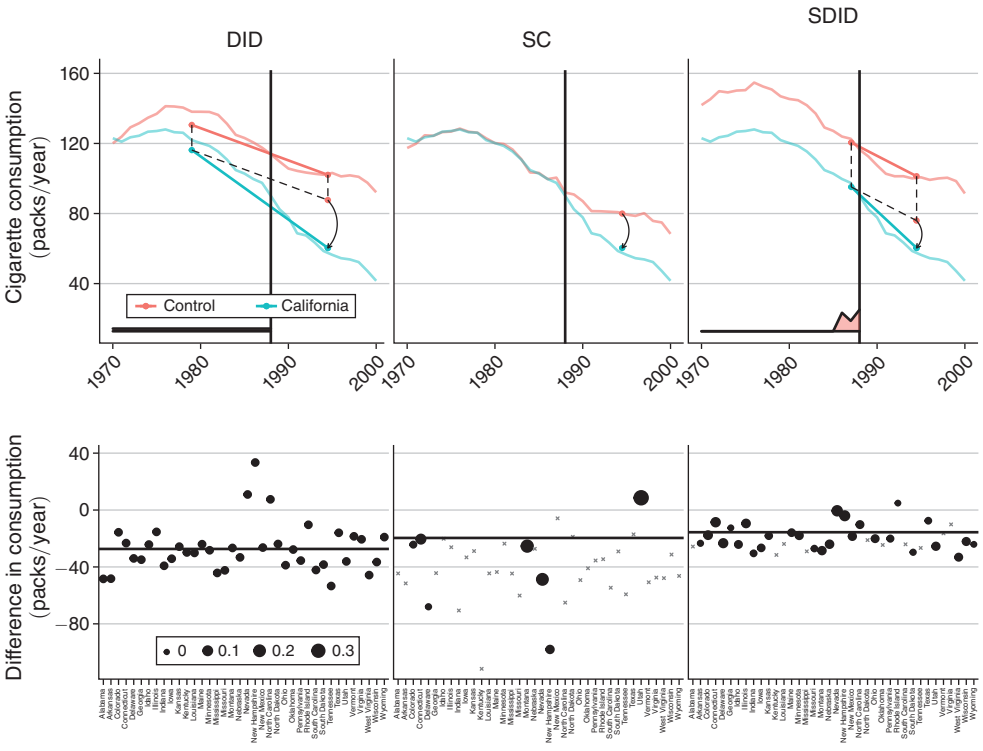

FIGURE 1. A COMPARISON BETWEEN DID, SC, AND SDID ESTIMATES FOR THE EFFECT OF CALIFORNIA PROPOSITION 99 ON PER-CAPITA ANNUAL CIGARETTE CONSUMPTION (IN PACKS/YEAR)

Notes: In the first row, we show trends in consumption over time for California and the relevant weighted average of control states, with the weights used to average pretreatment time periods at the bottom of the graphs. The estimated effect is indicated by an arrow. In the second row, we show the state-by-state adjusted outcome difference $\hat{\delta}_{tr} - \hat{\delta}_i$ as specified in (7) and (8), with the weights $\hat{\omega}_i$ indicated by dot size and the weighted average of these differences: the estimated effect—indicated by a horizontal line. States are ordered alphabetically. Observations with zero weight are denoted by an $\times$ symbol.

II. Placebo Studies

So far, we have relied on conceptual arguments to make the claim that SDID inherits good robustness properties from both traditional DID and SC methods, and shows promise as a method that can be used in settings where either DID and SC would traditionally be used. The goal of this section is to see how these claims play out in realistic empirical settings. To this end, we consider two carefully crafted simulation studies, calibrated to datasets representative of those typically used for panel data studies. The first simulation study mimics settings where DID would be used in practice (Section IIA), while the second mimics settings suited to SC (Section IIB). Not only do we base the outcome model of our simulation study on real datasets, we further ensure that the treatment assignment process is realistic by seeking to emulate the distribution of real policy initiatives. To be specific, in Section IIA, we consider a panel of US states. We estimate several alternative treatment assignment models to create the hypothetical treatments, where the models are based on the state laws related to minimum wages, abortion, or gun rights.

In order to run such a simulation study, we first need to commit to an econometric specification that can be used to assess the accuracy of each method. Here, we work with the following latent factor model (also referred to as an "interactive fixed-effects model" in Xu 2017; see also Athey et al. 2021),

$$Y_{it} = \boldsymbol{\gamma}_i \boldsymbol{\upsilon}_t^\top + \tau W_{it} + \varepsilon_{it}, \tag{9}$$

where $\boldsymbol{\gamma}_i$ is a vector of latent unit factors of dimension R, and $\boldsymbol{\upsilon}_t$ is a vector of latent time factors of dimension R. In matrix form, this can be written

$$\mathbf{Y} = \mathbf{L} + \tau \mathbf{W} + \mathbf{E} \qquad \text{where} \quad \mathbf{L} = \boldsymbol{\Gamma} \boldsymbol{\Upsilon}^\top. \tag{10}$$

We refer to $\mathbf{E}$ as the idiosyncratic component or error matrix, and to $\mathbf{L}$ as the systematic component. We assume that the conditional expectation of the error matrix $\mathbf{E}$ given the assignment matrix $\mathbf{W}$ and the systematic component $\mathbf{L}$ is zero. That is, the treatment assignment cannot depend on $\mathbf{E}$. However, the treatment assignment may in general depend on the systematic component $\mathbf{L}$ (i.e., we do not take $\mathbf{W}$ to be randomized). We assume that $\mathbf{E}_i$ is independent of $\mathbf{E}_{i'}$ for each pair of units i, i', but we allow for correlation across time periods within a unit. Our goal is to estimate the treatment effect τ.

The model (10) captures several qualitative challenges that have received considerable attention in the recent panel data literature. When the matrix $\mathbf{L}$ takes on an additive form, i.e., $L_{it} = \alpha_i + \beta_t$, then the DID regression will consistently recover τ. Allowing for interactions in $\mathbf{L}$ is a natural way to generalize the fixed-effects specification and discuss inference in settings where DID is misspecified (Bai 2009; Moon and Weidner 2015, 2017). In our formal results given in Section III, we show how, despite not explicitly fitting the model (10), SDID can consistently estimate τ in this design under reasonable conditions. Finally, accounting for correlation over time within observations of the same unit is widely considered to be an important ingredient to credible inference using panel data (Angrist and Pischke 2008; Bertrand, Duflo, and Mullainathan 2004).

In our experiments, we compare DID, SC, SDID, and DIFP, all implemented exactly as in Section I. We also compare these four estimators to an alternative that estimates τ by directly fitting both $\mathbf{L}$ and τ in (10); specifically, we consider the MC estimator recommended in Athey et al. (2021) that uses nuclear norm penalization to regularize its estimate of $\mathbf{L}$. In the remainder of this section, we focus on comparing the bias and root-mean-squared error (RMSE) of the estimator. We discuss questions around inference and coverage in Section IV.

A. *Current Population Survey Placebo Study*

Our first set of simulation experiments revisits the landmark placebo study of Bertrand, Duflo, and Mullainathan (2004) using the Current Population Survey (CPS). The main goal of Bertrand, Duflo, and Mullainathan (2004) was to study the behavior of different standard error estimators for DID. To do so, they randomly assigned a subset of states in the CPS dataset to a placebo treatment and the rest to the control group, and examined how well different approaches to inference for DID

estimators covered the true treatment effect of zero. Their main finding was that only methods that were robust to serial correlation of repeated observations for a given unit (e.g., methods that clustered observations by unit) attained valid coverage.

We modify the placebo analyses in Bertrand, Duflo, and Mullainathan (2004) in two ways. First, we no longer assigned exposed states completely at random, and instead use a nonuniform assignment mechanism that is inspired by different policy choices actually made by different states. Using a nonuniformly random assignment is important because it allows us to differentiate between various estimators in ways that completely random assignment would not. Under completely random assignment, a number of methods, including DID, perform well because the presence of $\mathbf{L}$ in (10) introduces zero bias. In contrast, with a nonuniform random assignment (i.e., treatment assignment is correlated with systematic effects), methods that do not account for the presence of $\mathbf{L}$ will be biased. Second, we simulate values for the outcomes based on a model estimated on the CPS data, in order to have more control over the data generating process.

The Data Generating Process.—For the first set of simulations we use as the starting point data on wages for women with positive wages in the March outgoing rotation groups in the CPS for the years 1979 to 2019. We first transform these by taking logarithms and then average them by state/year cells (we use data from National Bureau of Economic Research). Our simulation design has two components, an outcome model and an assignment model. We generate outcomes via a simulation that seeks to capture the behavior of the average by state/year of the logarithm of wages for those with positive hours worked in the CPS data as in Bertrand, Duflo, and Mullainathan (2004). Specifically, we simulate data using the model (10), where the rows $\mathbf{E}_i$ of $\mathbf{E}$ have a multivariate Gaussian distribution $\mathbf{E}_i \sim \mathcal{N}(0, \Sigma)$, and we choose both $\mathbf{L}$ and Σ to fit the CPS data as follows. First, we fit a rank four factor model for $\mathbf{L}$:

$$\mathbf{L} \coloneqq \underset{L:\mathrm{rank}(L)=4}{\arg\min} \sum_{it} \left(Y^*_{it} - L_{it}\right)^2, \tag{11}$$

where Y^*_{it} denotes the true state/year average of log wage in the CPS data. We then estimate Σ by fitting an AR(2) model to the residuals of $Y^*_{it} - L_{it}$. For purpose of interpretation, we further decompose the systematic component $\mathbf{L}$ into an additive (fixed effects) term $\mathbf{F}$ and an interactive term $\mathbf{M}$, with

$$\begin{aligned} F_{it} &= \alpha_i + \beta_t = \frac{1}{T}\sum_{l=1}^{T} L_{il} + \frac{1}{N}\sum_{j=1}^{N} L_{jt} - \frac{1}{NT}\sum_{it} L_{it}, \\ M_{it} &= L_{it} - F_{it}. \end{aligned} \tag{12}$$

This decomposition of $\mathbf{L}$ into an additive two-way fixed effect component $\mathbf{F}$ and an interactive component $\mathbf{M}$ enables us to study the sensitivity of different estimators to the presence of different types of systematic effects.

Next we discuss generation of the treatment assignment. Here, we are designing a "null effect" study, meaning that treatment has no effect on the outcomes and all methods should estimate zero. However, to make this more challenging, we choose

the treated units so that the assignment mechanism is correlated with the systematic component $\mathbf{L}$. We set $W_{it} = D_i \mathbf{1}_{t>T_0}$, where D_i is a binary exposure indicator generated as

$$D_i|\mathbf{E}_i, \alpha_i, \mathbf{M}_i \sim \text{Bernoulli}(\pi_i), \tag{13}$$

$$\pi_i = \pi(\alpha_i, \mathbf{M}_i; \phi) = \frac{\exp(\phi_\alpha \alpha_i + \phi_M \mathbf{M}_i)}{1 + \exp(\phi_\alpha \alpha_i + \phi_M \mathbf{M}_i)}.$$

In particular, the distribution of D_i may depend on α_i and $\mathbf{M}_i$; however, D_i is independent of $\mathbf{E}_i$, i.e., the assignment is strictly exogenous.[6] To construct probabilities $\{\pi_i\}$ for this assignment model, we choose ϕ as the coefficient estimates from a logistic regression of an observed binary characteristic of the state D_i on $\mathbf{M}_i$ and α_i. We consider three different choices for D_i, relating to minimum wage laws, abortion rights, and gun control laws.[7] As a result, we get assignment probability models that reflect actual differences across states with respect to important economic variables. In practice the α_i and $\mathbf{M}_i$ that we construct predict a sizable part of variation in D_i, with R^2 varying from 15 percent to 30 percent.

Simulation Results.—Table 2 compares the performance of the four aforementioned estimators in the simulation design described above. We consider various choices for the number of treated units and the treatment assignment distribution. Furthermore, we also consider settings where we drop various components of the outcome-generating process, such as the fixed effects $\mathbf{F}$ or the interactive component $\mathbf{M}$, or set the noise correlation matrix Σ to be diagonal. In the baseline simulation design (the first row of Table 2) these components have the following sizes: $\|\mathbf{F}\|_F/\sqrt{NT} = 0.992$, $\|\mathbf{M}\|_F/\sqrt{NT} = 0.100$, and $\sqrt{\text{tr}(\Sigma)/T} = 0.098$. The covariance matrix Σ is based on an AR(2) process with autoregressive coefficients $(\rho_{-1}, \rho_{-2}) = (0.01, -0.06)$.

At a high level, we find that SDID has excellent performance relative to the benchmarks—both in terms of bias and RMSE. This holds in the baseline simulation design and over a number of other designs where we vary the treatment assignment (from being based on minimum wage laws to gun laws, abortion laws, or completely random), the outcome (from average of log wages to average hours and unemployment rate), and the maximal number of treated units (from ten to one) and the number of exposed periods (from ten to one). We find that when the treatment assignment is uniformly random, all methods are essentially unbiased, but SDID is more precise. Meanwhile, when the treatment assignment is not uniformly random, SDID is particularly successful at mitigating bias while keeping variance in check.

In the second panel of Table 2 we provide some additional insights into the superior performance of the SDID estimator by sequentially dropping some of the components of the model that generates the potential outcomes. If we drop the interactive component $\mathbf{M}$ from the outcome model ("No $\mathbf{M}$"), so that the fixed

[6] In the simulations below, we restrict the maximal number of treated units (either to ten or one). To achieve this, we first sample D_i independently and accept the results if the number of treated units satisfies the constraint. If it does not, then we choose the maximal allowed number of treated units from those selected in the first step uniformly at random.

[7] See the Appendix for details.

TABLE 2

	RMSE					Bias				
	SDID	SC	DID	MC	DIFP	SDID	SC	DID	MC	DIFP
1. Baseline	**0.28**	0.37	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
Outcome model										
2. No corr	**0.28**	0.38	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
3. No **M**	0.16	0.18	**0.14**	**0.14**	0.16	0.01	0.04	0.01	0.01	0.01
3. No **F**	0.28	**0.23**	0.49	0.35	0.32	0.10	0.04	0.21	0.15	0.07
4. Only noise	0.16	**0.14**	**0.14**	**0.14**	0.16	0.01	0.01	0.01	0.01	0.01
5. No noise	0.06	0.17	0.47	**0.04**	0.11	0.05	0.04	0.20	0.00	0.01
Assignment process										
6. Gun law	**0.26**	0.27	0.47	0.36	0.30	0.08	−0.03	0.15	0.15	0.09
7. Abortion	**0.23**	0.31	0.45	0.31	0.27	0.04	0.16	0.03	0.02	0.01
8. Random	**0.24**	0.25	0.44	0.31	0.27	0.01	−0.01	0.02	0.01	−0.00
Outcome variable										
9. Hours	1.90	2.03	2.06	**1.85**	1.97	1.12	−0.49	0.85	1.00	1.00
10. U-rate	**2.25**	2.31	3.91	2.96	2.30	1.77	1.73	3.60	2.63	1.69
Assignment block size										
11. $T_{post} = 1$	**0.50**	0.59	0.70	0.51	0.54	0.20	0.17	0.38	0.21	0.12
12. $N_{tr} = 1$	**0.63**	0.73	1.26	0.81	0.83	0.03	0.15	0.11	0.05	−0.02
13. $T_{post} = N_{tr} = 1$	1.12	1.24	1.52	**1.07**	1.16	0.14	0.24	0.33	0.16	0.11

Notes: Simulation results for CPS data. The baseline case uses state minimum wage laws to simulate treatment assignment, and generates outcomes using the full data-generating process described in Section IIA, with $T_{post} = 10$ posttreatment periods and at most $N_{tr} = 10$ treatment states. In subsequent settings, we omit parts of the data-generating process (rows 2–6), consider different distributions for the treatment exposure variable D_i (rows 7–9) and different distributions for the outcome variable (rows 10 and 11), and vary the number of treated cells (rows 12–14). The full dataset has $N = 50$, $T = 40$, and outcomes are normalized to have mean zero and unit variance. All results are based on 1,000 simulation replications and are multiplied by ten for readability.

effect specification is correct, the DID estimator performs best (alongside MC). In contrast, if we drop the fixed effects component ("No **F**") but keep the interactive component, the SC estimator does best. If we drop both parts of the systematic component, and there is only noise, the superiority of the SDID estimator vanishes and all estimators are essentially equivalent. On the other hand, if we remove the noise component so that there is only signal, the increased flexibility of the SDID estimator allows it (alongside MC) to outperform the SC and DID estimators dramatically.

Next, we focus on two designs of interest: one with the assignment probability model based on parameters estimated in the minimum wage law model and one where the treatment exposure D_i is assigned uniformly at random. Figure 2 shows the errors of the DID, SC, and SDID estimators in both settings, and reinforces our observations above. When assignment is not uniformly random, the distribution of the DID errors is visibly off-center, showing the bias of the estimator. In contrast, the errors from SDID are nearly centered. Meanwhile, when treatment assignment is uniformly random, both estimators are centered but the errors of DID are more spread out. We note that the right panel of Figure 2 is closely related to the simulation specification of Bertrand, Duflo, and Mullainathan (2004). From this perspective, Bertrand, Duflo, and Mullainathan (2004) correctly argue that the error distribution of DID is centered, and that the error scale can accurately be recovered

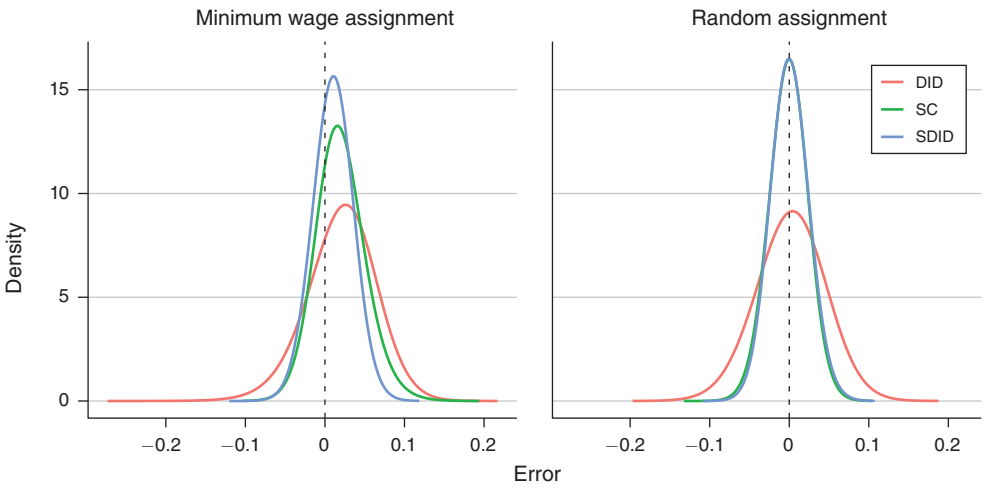

FIGURE 2. DISTRIBUTION OF THE ERRORS OF SDID, SC, AND DID IN THE SETTING OF THE "BASELINE" (I.E., WITH MINIMUM WAGE) AND RANDOM ASSIGNMENT ROWS OF TABLE 2

using appropriate robust estimators. Here, however, we go further and show that this noise can be substantially reduced by using an estimator like SDID that can exploit predictable variation by matching on pre-exposure trends.

Finally, we note that Figure 2 shows that the error distribution of SDID is nearly unbiased and Gaussian in both designs, thus suggesting that it should be possible to use $\hat{\tau}^{sdid}$ as the basis for valid inference. We postpone a discussion of confidence intervals until Section IV, where we consider various strategies for inference based on SDID and show that they attain good coverage here.

B. *Penn World Table Placebo Study*

The simulation based on the CPS is a natural benchmark for applications that traditionally rely on DID-type methods to estimate the policy effects. In contrast, SC methods are often used in applications where units tend to be more heterogeneous and are observed over a longer timespan as in, e.g., Abadie, Diamond, and Hainmueller (2015). To investigate the behavior of SDID in this type of setting, we propose a second set of simulations based on the Penn World Table (Feenstra, Inklaar, and Timmer 2015). This dataset contains observations on annual real GDP for $N = 111$ countries for $T = 48$ consecutive years, starting from 1959; we end the dataset in 2007 because we do not want the treatment period to coincide with the Great Recession. We construct the outcome and the assignment model following the same procedure outlined in the previous subsection. We select log(realGDP) as the primary outcome. As with the CPS dataset, the two-way fixed effects explain most of the variation; however, the interactive component plays a larger role in determining outcomes for this dataset than for the CPS data. We again derive treatment assignment via an exposure variable D_i, and consider both a uniformly random distribution for D_i as well as two nonuniform ones based on predicting Penn World Table indicators of democracy and education respectively.

TABLE 3

	RMSE					Bias				
	SDID	SC	DID	MC	DIFP	SDID	SC	DID	MC	DIFP
Democracy	**0.31**	0.38	1.97	0.58	0.39	−0.05	−0.04	1.75	0.43	−0.07
Education	**0.30**	0.53	1.72	0.49	0.39	−0.03	0.25	1.62	0.40	−0.05
Random	**0.37**	0.46	1.29	0.63	0.45	−0.02	−0.11	−0.06	−0.04	−0.04

Notes: Simulation results based on the Penn World Table dataset. We use $\log(GDP)$ as the outcome, with $N_{tr} = 10$ out of $N = 111$ treatment countries, and $T_{post} = 10$ out of $T = 48$ treatment periods. In the first two rows we consider treatment assignment distributions based on democracy status and education metrics, while in the last row the treatment is assigned completely at random. All results are based on 1,000 simulations and multiplied by ten for readability.

Results of the simulation study are presented in Table 3. At a high level, these results mirror the ones above: SDID again performs well in terms of both bias and RMSE and across all simulation settings dominates the other estimators. In particular, SDID is nearly unbiased, which is important for constructing confidence intervals with accurate coverage rates. The main difference between Tables 2 and 3 is that DID does substantially worse here relative to SC than before. This appears to be due to the presence of a stronger interactive component in the Penn World Table dataset, and is in line with the empirical practice of preferring SC over DID in settings of this type. We again defer a discussion of inference to Section IV.

III. Formal Results

In this section we discuss the formal results. For the remainder of the paper, we assume that the data generating process follows a generalization of the latent factor model (10),

$$\mathbf{Y} = \mathbf{L} + \mathbf{W} \circ \boldsymbol{\tau} + \mathbf{E}, \quad \text{where} \quad (\mathbf{W} \circ \boldsymbol{\tau})_{it} = \mathbf{W}_{it}\boldsymbol{\tau}_{it}. \tag{14}$$

The model allows for heterogeneity in treatment effects τ_{it}, as in de Chaisemartin and D'Haultfœuille (2020). As above, we assume block assignment $W_{it} = \mathbf{1}(\{i > N_{co}, t > T_{pre}\})$, where the subscript "*co*" stands for control group, "*tr*" stands for treatment group, "*pre*" stands for pretreatment, and "*post*" stands for posttreatment. It is useful to characterize the systematic component $\mathbf{L}$ as a factor model $\mathbf{L} = \boldsymbol{\Gamma}\boldsymbol{\Upsilon}^\top$ as in (10), where we define factors $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}^{1/2}$ and $\boldsymbol{\Upsilon}^\top = \mathbf{D}^{1/2}\mathbf{V}^\top$ in terms of the singular value decomposition $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Our target estimand is the average treatment effect for the treated units during the periods they were treated, which under block assignment is

$$\tau = \frac{1}{N_{tr}T_{post}} \sum_{i=N_{co}+1}^{N} \sum_{t=T_{pre}+1}^{T} \boldsymbol{\tau}_{it}. \tag{15}$$

For notational convenience, we partition the matrix $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{co,pre} & \mathbf{Y}_{co,post} \\ \mathbf{Y}_{tr,pre} & \mathbf{Y}_{tr,post} \end{pmatrix},$$

with $\mathbf{Y}_{co,pre}$, an $N_{co} \times T_{pre}$ matrix; $\mathbf{Y}_{co,post}$, an $N_{co} \times T_{post}$ matrix; $\mathbf{Y}_{tr,pre}$, an $N_{tr} \times T_{pre}$ matrix; and $\mathbf{Y}_{tr,post}$, an $N_{tr} \times T_{post}$ matrix; and similar for $\mathbf{L}$, $\mathbf{W}$, τ, and $\mathbf{E}$. Throughout our analysis, we will assume that the errors $\mathbf{E}_{i\cdot}$ are homoskedastic across units (but not across time), i.e., that $\text{var}[\mathbf{E}_{i\cdot}] = \Sigma \in \mathbb{R}^{T \times T}$ for all units $i = 1, \ldots, n$. We partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{pre,pre} & \Sigma_{pre,post} \\ \Sigma_{post,pre} & \Sigma_{post,post} \end{pmatrix}.$$

Given this setting, we are interested in guarantees on how accurately SDID can recover τ.

A simple, intuitively appealing approach to estimating τ in (14) is to directly fit both $\mathbf{L}$ and τ via methods for low-rank matrix estimation, and several variants of this approach have been proposed in the literature (e.g., Athey et al. 2021, Bai 2009, Xu 2017, Agarwal et al. 2019). However, our main interest is in τ and not in $\mathbf{L}$, and so one might suspect that approaches that provide consistent estimation of $\mathbf{L}$ may rely on assumptions that are stronger than what is necessary for consistent estimation of τ.

SC methods address confounding bias without explicitly estimating $\mathbf{L}$ in (14). Instead, they take an indirect approach more akin to balancing as in Zubizarreta (2015) and Athey, Imbens, and Wager (2018). Recall that the SC weights $\hat{\omega}^{sc}$ seek to balance out the preintervention trends in $\mathbf{Y}$. Qualitatively, one might hope that doing so also leads us to balance out the unit factors $\mathbf{\Gamma}$ from (10), rendering $\sum_{i=N_{co}+1}^{N} \hat{\omega}_i^{sc} \mathbf{\Gamma}_{i\cdot} - \sum_{i=1}^{N_{co}} \hat{\omega}_i^{sc} \mathbf{\Gamma}_{i\cdot} \approx 0$. Abadie, Diamond, and Hainmueller (2010) provide some arguments for why this should be the case, and our formal analysis outlines a further set of conditions under which this type of phenomenon holds. Then, if $\hat{\omega}^{sc}$ in fact succeeds in balancing out the factors in $\mathbf{\Gamma}$, the SC estimator can be approximated as $\hat{\tau}^{sc} \approx \tau + \sum_{i=1}^{N} (2W_i - 1)\, \hat{\omega}_i^{sc} \bar{\varepsilon}_i$ with $\bar{\varepsilon}_i = T_{post}^{-1} \sum_{t=T_{pre}+1}^{T} \varepsilon_{it}$; in words, SC weighting has succeeded in removing the bias associated with the systematic component $\mathbf{L}$ and in delivering a nearly unbiased estimate of τ.

Much like the SC estimator, the SDID estimator seeks to recover τ in (14) by reweighting to remove the bias associated with $\mathbf{L}$. However, the SDID estimator takes a two-pronged approach. First, instead of only making use of unit weights $\hat{\omega}$ that can be used to balance out $\mathbf{\Gamma}$, the estimator also incorporates time weights $\hat{\lambda}$ that seek to balance out $\mathbf{\Upsilon}$. This provides a type of double robustness property, whereby if one of the balancing approaches is effective, the dependence on $\mathbf{L}$ is approximately removed. Second, the use of two-way fixed effects in (1) and intercept terms in (4) and (6) makes the SDID estimator invariant to additive shocks to any row or column; i.e., if we modify $\mathbf{L}_{it} \leftarrow \mathbf{L}_{it} + \alpha_i + \beta_t$ for any choices α_i and β_t the estimator $\hat{\tau}^{sdid}$ remains unchanged. The estimator shares this invariance property with DID (but not SC).[8]

The goal of our formal analysis is to understand how and when the SDID weights succeed in removing the bias due to $\mathbf{L}$. As discussed below, this requires assumptions

[8] More specifically, as suggested by (3), SC is invariant to shifts in β_t but not α_i. In this context, we also note that the DIFP estimator proposed by Doudchenko and Imbens (2016) and Ferman and Pinto (2019) that center each unit's trajectory before applying the SC method is also invariant to shifts in α_i.

on the signal to noise ratio. The assumptions require that $\mathbf{E}$ does not incorporate too much serial correlation within units, so that we can attribute persistent patterns in $\mathbf{Y}$ to patterns in $\mathbf{L}$; furthermore, $\mathbf{\Gamma}$ should be stable over time, particularly through the treatment periods. Of course, these are nontrivial assumptions. However, as discussed further in Section V, they are considerably weaker than what is required in results of Bai (2009) or Moon and Weidner (2015, 2017) for methods that require explicitly estimating $\mathbf{L}$ in (14). Furthermore, these assumption are aligned with standard practice in the literature; for example, we can assess the claim that we balance all components of $\mathbf{\Gamma}$ by examining the extent to which the method succeeds in balancing preintervention periods. Historical context may be needed to justify the assumption that there were no other shocks disproportionately affecting the treatment units at the time of the treatment.

A. *Weighted Double-Differencing Estimators*

We introduced the SDID estimator (1) as the solution to a weighted two-way fixed effects regression. For the purpose of our formal results, however, it is convenient to work with the alternative characterization described in equation (16). For any weights $\omega \in \Omega$ and $\lambda \in \Lambda$, we can define a weighted double-differencing estimator[9]

$$(16)\quad \hat{\tau}(\omega, \lambda) = \omega_{tr}^\top \mathbf{Y}_{tr,post}\lambda_{post} - \omega_{co}^\top \mathbf{Y}_{co,post}\lambda_{post} - \omega_{tr}^\top \mathbf{Y}_{tr,pre}\lambda_{pre} + \omega_{co}^\top \mathbf{Y}_{co,pre}\lambda_{pre}.$$

One can verify that the basic DID estimator is of the form (16), with constant weights $\omega_{tr} = 1/N_{tr}$, etc. The proposed SDID estimator (1) can also be written as (16), but now with weights $\hat{\omega}^{sdid}$ and $\hat{\lambda}^{sdid}$ solving (4) and (6) respectively. When there is no risk of ambiguity, we will omit the SDID superscript from the weights and simply write $\hat{\omega}$ and $\hat{\lambda}$.

Now, note that for any choice of weights $\omega \in \Omega$ and $\lambda \in \Lambda$, we have $\omega_{tr} \in \mathbb{R}^{N_{tr}}$ and $\lambda_{post} \in \mathbb{R}^{T_{post}}$ with all elements equal to $1/N_{tr}$ and $1/T_{post}$ respectively, and so $\omega_{tr}^\top \boldsymbol{\tau}_{tr,post}\lambda_{post} = \tau$. Thus, we can decompose the error of any weighted double-differencing estimator with weights satisfying these conditions as the sum of a bias and a noise component:

$$\begin{aligned}(17)\quad &\hat{\tau}(\omega, \lambda) - \tau \\ &= \underbrace{\omega_{tr}^\top \mathbf{L}_{tr,post}\lambda_{post} - \omega_{co}^\top \mathbf{L}_{co,post}\lambda_{post} - \omega_{tr}^\top \mathbf{L}_{tr,pre}\lambda_{pre} + \omega_{co}^\top \mathbf{L}_{co,pre}\lambda_{pre}}_{\text{bias } B(\omega,\lambda)} \\ &\quad + \underbrace{\omega_{tr}^\top \mathbf{E}_{tr,post}\lambda_{post} - \omega_{co}^\top \mathbf{E}_{co,post}\lambda_{post} - \omega_{tr}^\top \mathbf{E}_{tr,pre}\lambda_{pre} + \omega_{co}^\top \mathbf{E}_{co,pre}\lambda_{pre}}_{\text{noise } \varepsilon(\omega,\lambda)}.\end{aligned}$$

[9] This weighted double-differencing structure plays a key role in understanding the behavior of SDID. As discussed further in Section V, despite relying on a different motivation, certain specifications of the recently proposed "augmented synthetic control" method of Ben-Michael, Feller, and Rothstein (2018) also result in a weighted double-differencing estimator.

In order to characterize the distribution of $\hat{\tau}^{sdid} - \tau$, it thus remains to carry out two tasks. First, we need to understand the scale of the errors $B(\omega, \lambda)$ and $\varepsilon(\omega, \lambda)$, and second, we need to understand how data adaptivity of the weights $\hat{\omega}$ and $\hat{\lambda}$ affects the situation.

B. *Oracle and Adaptive Synthetic Control Weights*

To address the adaptivity of the SDID weights $\hat{\omega}$ and $\hat{\lambda}$ chosen via (4) and (6), we construct alternative "oracle" weights that have similar properties to $\hat{\omega}$ and $\hat{\lambda}$ in terms of eliminating bias due to $\mathbf{L}$, but are deterministic. We can then further decompose the error of $\hat{\tau}^{sdid}$ into the error of a weighted double-differencing estimator with the oracle weights and the difference between the oracle and feasible estimators. Under appropriate conditions, we find the latter term negligible relative to the error of the oracle estimator, opening the door to a simple asymptotic characterization of the error distribution of $\hat{\tau}^{sdid}$.

We define such oracle weights $\tilde{\omega}$ and $\tilde{\lambda}$ by minimizing the expectation of the objective functions $\ell_{unit}(\cdot)$ and $\ell_{time}(\cdot)$ used in (4) and (6) respectively, and set

$$(18)\quad (\tilde{\omega}_0, \tilde{\omega}) = \operatorname*{argmin}_{\omega_0 \in \mathbb{R}, \omega \in \Omega} E\big[\ell_{unit}(\omega_0, \omega)\big], \qquad \big(\tilde{\lambda}_0, \tilde{\lambda}\big) = \operatorname*{argmin}_{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda} E\big[\ell_{time}(\lambda_0, \lambda)\big].$$

In the case of our model (14) these weights admit a simplified characterization

$$(19)\quad (\tilde{\omega}_0, \tilde{\omega}) = \operatorname*{argmin}_{\omega_0 \in \mathbb{R}, \omega \in \Omega} ||\omega_0 + \omega_{co}^\top \mathbf{L}_{co,pre} - \omega_{tr}^\top \mathbf{L}_{tr,pre}||_2^2 + \Big(\operatorname{tr}\big(\Sigma_{pre,pre}\big) + \zeta^2 T_{pre}\Big) ||\omega||_2^2,$$

$$(20)\quad \big(\tilde{\lambda}_0, \tilde{\lambda}\big) = \operatorname*{argmin}_{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda} ||\lambda_0 + \mathbf{L}_{co,pre}\lambda_{pre} - \mathbf{L}_{co,post}\lambda_{post}||_2^2 + ||\tilde{\Sigma}\lambda||_2^2,$$

where

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_{pre,pre} & -\Sigma_{pre,post} \\ -\Sigma_{post,pre} & \Sigma_{post,post} \end{pmatrix}.$$

The error of the SDID estimator can now be decomposed as follows,

$$(21)\quad \hat{\tau}^{sdid} - \tau = \underbrace{\varepsilon\big(\tilde{\omega}, \tilde{\lambda}\big)}_{\text{oracle noise}} + \underbrace{B\big(\tilde{\omega}, \tilde{\lambda}\big)}_{\text{oracle confounding bias}} + \underbrace{\hat{\tau}\big(\hat{\omega}, \hat{\lambda}\big) - \hat{\tau}\big(\tilde{\omega}, \tilde{\lambda}\big)}_{\text{deviation from oracle}},$$

and our task is to characterize all three terms.

First, the oracle noise term tends to be small when the weights are not too concentrated, i.e., when $\|\tilde{\omega}\|_2$ and $\|\tilde{\lambda}\|_2$ are small, and we have a sufficient number of exposed units and time periods. In the case with $\Sigma = \sigma^2 I_{T \times T}$, i.e., without any cross-observation correlations, we note that $\operatorname{var}[\varepsilon(\tilde{\omega}, \tilde{\lambda})] = \sigma^2(N_{tr}^{-1} + \|\tilde{\omega}\|_2^2)(T_{post}^{-1} + \|\tilde{\lambda}\|_2^2)$. When we move to our asymptotic analysis below, we work under assumptions that make this oracle noise term dominant relative to the other error terms in (21).

Second, the oracle confounding bias will be small either when the pre-exposure oracle row regression fits well and generalizes to the exposed rows, i.e., $\tilde{\omega}_0 + \tilde{\omega}_{co}^\top \mathbf{L}_{co,pre} \approx \tilde{\omega}_{tr}^\top \mathbf{L}_{tr,pre}$ and $\tilde{\omega}_0 + \tilde{\omega}_{co}^\top \mathbf{L}_{co,post} \approx \tilde{\omega}_{tr}^\top \mathbf{L}_{tr,post}$, or when the unexposed oracle column regression fits well and generalizes to the exposed columns, $\tilde{\lambda}_0 + \mathbf{L}_{co,pre}\tilde{\lambda}_{pre} \approx \mathbf{L}_{co,post}\tilde{\lambda}_{post}$ and $\tilde{\lambda}_0 + \mathbf{L}_{tr,pre}\tilde{\lambda}_{pre} \approx \mathbf{L}_{tr,post}\tilde{\lambda}_{post}$. Moreover, even if neither model generalizes sufficiently well on its own, it suffices for one model to predict the generalization error of the other:

$$\begin{aligned} B(\omega,\lambda) &= \left(\omega_{tr}^\top \mathbf{L}_{tr,post} - \omega_{co}^\top \mathbf{L}_{co,post}\right)\lambda_{post} - \left(\omega_{tr}^\top \mathbf{L}_{tr,pre} - \omega_{co}^\top \mathbf{L}_{co,pre}\right)\lambda_{pre} \\ &= \omega_{tr}^\top\left(\mathbf{L}_{tr,post}\lambda_{post} - \mathbf{L}_{tr,pre}\lambda_{pre}\right) - \omega_{co}^\top\left(\mathbf{L}_{co,post}\lambda_{post} - \mathbf{L}_{co,pre}\lambda_{pre}\right). \end{aligned}$$

The upshot is even if one of the sets of weights fails to remove the bias from the presence of $\mathbf{L}$, the combination of weights $\tilde{\omega}$ and $\tilde{\lambda}$ can compensate for such failures. This double robustness property is similar to that of the augmented inverse probability weighting estimator, whereby one can trade off between accurate estimates of the outcome and treatment assignment models (Ben-Michael, Feller, and Rothstein 2018; Scharfstein, Rotnitzky, and Robins 1999).

We note that although poor fit in the oracle regressions on the unexposed rows and columns of $\mathbf{L}$ will often be indicated by a poor fit in the realized regressions on the unexposed rows and columns of $\mathbf{Y}$, the assumption that one of these regressions generalizes to exposed rows or columns is an identification assumption without clear testable implications. It is essentially an assumption of no unexplained confounding: any exceptional behavior of the exposed observations, whether due to exposure or not, can be ascribed to it.

Third, our core theoretical claim, formalized in our asymptotic analysis, is that the SDID estimator will be close to the oracle when the oracle unit and time weights look promising on their respective training sets, i.e, when $\tilde{\omega}_0 + \tilde{\omega}_{co}^\top \mathbf{L}_{co,pre} \approx \tilde{\omega}_{tr}^\top \mathbf{L}_{tr,pre}$ and $\|\tilde{\omega}\|_2$ is not too large and $\tilde{\lambda}_0 + \mathbf{L}_{co,pre}\tilde{\lambda}_{pre} \approx \mathbf{L}_{co,post}\tilde{\lambda}_{post}$ and $\|\tilde{\lambda}\|_2$ is not too large. Although the details differ, as described above these qualitative properties are also criteria for accuracy of the oracle estimator itself.

Finally, we comment briefly on the behavior of the oracle time weights $\tilde{\lambda}$ in the presence of autocorrelation over time. When Σ is not diagonal, the effective regularization term in (20) does not shrink $\tilde{\lambda}_{pre}$ towards zero, but rather toward an autoregression vector

$$\psi = \underset{v \in \mathbb{R}^{T_{pre}}}{\arg\min} \left\| \tilde{\Sigma}\begin{pmatrix} v \\ \lambda_{post} \end{pmatrix} \right\| = \Sigma_{pre,pre}^{-1}\Sigma_{pre,post}\lambda_{post}. \tag{22}$$

Here λ_{post} is the T_{post}-component column vector with all elements equal to $1/T_{post}$ and ψ is the population regression coefficient in a regression of the average of the posttreatment errors on the pretreatment errors. In the absence of autocorrelation, ψ is zero, but when autocorrelation is present, shrinkage toward ψ reduces the variance of the SDID estimator—and enables us to gain precision over the basic DID estimator (2) even when the two-way fixed effects model is correctly specified. This explains some of the behavior noted in the simulations.

C. *Asymptotic Properties*

To carry out the analysis plan sketched above, we need to embed our problem into an asymptotic setting. First, we require the error matrix $\mathbf{E}$ to satisfy some regularity properties.

ASSUMPTION 1 (Properties of Errors): *The rows $\mathbf{E}_i$ of the noise matrix are independent and identically distributed Gaussian vectors and the eigenvalues of its covariance matrix Σ are bounded and bounded away from zero.*

Next, we spell out assumptions about the sample size. At a high level, we want the panel to be large (i.e., $N, T \to \infty$), and for the number of treated cells of the panel to grow to infinity but slower than the total panel size. We note in particular that we can accommodate sequences where one of T_{post} or N_{tr} is fixed, but not both.

ASSUMPTION 2 (Sample Sizes): *We consider a sequence of populations where*

(*i*) *the product $N_{tr} T_{post}$ goes to infinity, and both N_{co} and T_{pre} go to infinity,*

(*ii*) *the ratio T_{pre}/N_{co} is bounded and bounded away from zero,*

(*iii*) $N_{co}/(N_{tr} T_{post} \max(N_{tr}, T_{post}) \log^2(N_{co})) \to \infty$.

We also need to make assumptions about the spectrum of $\mathbf{L}$; in particular, $\mathbf{L}$ cannot have too many large singular values, although we allow for the possibility of many small singular values. A sufficient, but not necessary, condition for the assumption below is that the rank of $\mathbf{L}$ is less than $\sqrt{\min(T_{pre}, N_{co})}$. Notice that we do not assume any lower bounds for nonzero singular values of $\mathbf{L}$; in fact can accommodate arbitrarily many nonzero but very small singular values, much like, e.g., Belloni, Chernozhukov, and Hansen (2014) can accommodate arbitrarily many nonzero but very small signal coefficients in a high-dimensional inference problem. We need the $\sqrt{\min(T_{pre}, N_{co})}$th singular value of $\mathbf{L}_{co,pre}$ to be sufficiently small. Formally, we have the following result.

ASSUMPTION 3 (Properties of $\mathbf{L}$): *Letting $\sigma_1(\Gamma), \sigma_2(\Gamma), \ldots$ denote the singular values of the matrix Γ in decreasing order and R the largest integer less than $\sqrt{\min(T_{pre}, N_{co})}$,*

$$\sigma_R(\mathbf{L}_{co,pre})/R = o\Big(\min\Big\{N_{tr}^{-1/2}\log^{-1/2}(N_{co}), T_{post}^{-1/2}\log^{-1/2}(T_{pre})\Big\}\Big). \tag{23}$$

The last—and potentially most interesting—of our assumptions concerns the relation between the factor structure $\mathbf{L}$ and the assignment mechanism $\mathbf{W}$. At a high level, it plays the role of an identifying assumption, and guarantees that the oracle weights from (19) and (20) that are directly defined in terms of $\mathbf{L}$ are able to adequately cancel out $\mathbf{L}$ via the weighted double-differencing strategy. This requires that the optimization problems (19) and (20) accommodate reasonably dispersed

weights, and that the treated units and after periods not be too dissimilar from the control units and the before periods respectively.

ASSUMPTION 4 (Properties of Weights and $\mathbf{L}$): *The oracle unit weights $\tilde{\omega}$ satisfy*

$$\|\tilde{\omega}_{co}\|_2 = o\Big(\big[(N_{tr}T_{post})\log(N_{co})\big]^{-1/2}\Big) \tag{24}$$

and

$$\|\tilde{\omega}_0 + \tilde{\omega}_{co}^\top \mathbf{L}_{co,pre} - \tilde{\omega}_{tr}^\top \mathbf{L}_{tr,pre}\|_2 = o\Big(N_{co}^{1/4}\big(N_{tr}T_{post}\max(N_{co},T_{post})\big)^{-1/4}\log^{-1/2}(N_{co})\Big),$$

the oracle time weights $\tilde{\lambda}$ satisfy

$$\|\tilde{\lambda}_{pre} - \psi\|_2 = o\Big(\big[(N_{tr}T_{post})\log(N_{co})\big]^{-1/2}\Big) \tag{25}$$

and

$$\|\tilde{\lambda}_0 + \mathbf{L}_{co,pre}\tilde{\lambda}_{pre} - \mathbf{L}_{co,post}\tilde{\lambda}_{post}\|_2 = o\Big(N_{co}^{1/4}\big(N_{tr}T_{post}\big)^{-1/8}\Big),$$

and the oracle weights jointly satisfy

$$\tilde{\omega}_{tr}^\top \mathbf{L}_{tr,post}\tilde{\lambda}_{post} - \tilde{\omega}_{co}^\top \mathbf{L}_{co,post}\tilde{\lambda}_{post} - \tilde{\omega}_{tr}^\top \mathbf{L}_{tr,pre}\tilde{\lambda}_{pre} + \tilde{\omega}_{co}^\top \mathbf{L}_{co,pre}\tilde{\lambda}_{pre} = o\Big(\big(N_{tr}T_{post}\big)^{-1/2}\Big). \tag{26}$$

Assumptions 1–4 are substantially weaker than those used to establish asymptotic normality of comparable methods.[10] We do not require that double differencing alone removes the individual and time effects as the DID assumptions do. Furthermore, we do not require that unit comparisons alone are sufficient to remove the biases in comparisons between treated and control units as the SC assumptions do. Finally, we do not require a low rank factor model to be correctly specified, as is often assumed in the analysis of methods that estimate $\mathbf{L}$ explicitly (e.g., Bai 2009, Moon and Weidner 2015, 2017). Rather, we only need the combination of the three bias-reducing components in the SDID estimator, (i) double differencing, (ii) the unit weights, and (iii) the time weights, to reduce the bias to a sufficiently small level.

Our main formal result states that under these assumptions, our estimator is asymptotically normal. Furthermore, its asymptotic variance is optimal,

[10] In particular, note that our assumptions are satisfied in the well-specified two-way fixed effect setting model. Suppose we have $L_{it} = \alpha_i + \beta_t$ with uncorrelated and homoskedastic errors, and that the sample size restrictions in Assumption 2 are satisfied. Then Assumption 1 is automatically satisfied, and the rank condition on $\mathbf{L}$ from Assumption 3 is satisfied with $R = 2$. Next, we see that the oracle unit weights satisfy $\tilde{\omega}_{co,i} = 1/N_{co}$ so that $\|\tilde{\omega}\|_2 = 1/\sqrt{N_{co}}$, and the oracle time weights satisfy $\tilde{\lambda}_{pre,i} = 1/T_{pre}$ so that $\|\tilde{\lambda} - \psi\|_2 = 1/\sqrt{N_{co}}$. Thus if the restrictions on the rates at which the sample sizes increase in Assumption 2 are satisfied, then (24) and (25) are satisfied. Finally, the additive structure of $\mathbf{L}$ implies that, as long as the weights for the controls sum to one, $\tilde{\omega}_{tr}^\top \mathbf{L}_{tr,post}\tilde{\lambda}_{post} - \tilde{\omega}_{co}^\top \mathbf{L}_{co,post}\tilde{\lambda}_{post} = 0$, and $\tilde{\omega}_{tr}^\top \mathbf{L}_{tr,pre}\tilde{\lambda}_{pre} + \tilde{\omega}_{co}^\top \mathbf{L}_{co,pre}\tilde{\lambda}_{pre} = 0$, so that (26) is satisfied.

coinciding with the variance we would get if we knew $\mathbf{L}$ and Σ a priori and could therefore estimate τ by a simple average of τ_{it} plus unpredictable noise, $N_{tr}^{-1}\sum_{i=N_{co}+1}^{N}[T_{post}^{-1}\sum_{t=T_{pre}+1}^{T}(\boldsymbol{\tau}_{it}+\varepsilon_{it})-\mathbf{E}_{i,pre}\psi]$.

THEOREM 1: *Under the model (14) with* $\mathbf{L}$ *and* $\mathbf{W}$ *taken as fixed, suppose that we run the SDID estimator (1) with regularization parameter* ζ *satisfying* $(N_{tr}T_{post})^{1/2}\log(N_{co}) = o(\zeta^2)$. *Suppose moreover that Assumptions 1–4 hold. Then,*

$$\hat{\tau}^{sdid}-\tau=\frac{1}{N_{tr}}\sum_{i=N_{co}+1}^{N}\left(\frac{1}{T_{post}}\sum_{t=T_{pre}+1}^{T}\varepsilon_{it}-\mathbf{E}_{i,pre}\psi\right)+o_p\Big(\big(N_{tr}T_{post}\big)^{-1/2}\Big), \tag{27}$$

and consequently

$$\big(\hat{\tau}^{sdid}-\tau\big)/V_{\tau}^{1/2}\Rightarrow\mathcal{N}(0,1), \tag{28}$$

where

$$V_{\tau}=\frac{1}{N_{tr}}\mathrm{var}\left[\frac{1}{T_{post}}\sum_{t=T_{pre}+1}^{T}\varepsilon_{it}-\mathbf{E}_{i,pre}\psi\right].$$

Here V_{τ} *is on the order of* $1/(N_{tr}T_{post})$, i.e., $N_{tr}T_{post}V_{\tau}$ *is bounded and bounded away from zero.*

IV. Large-Sample Inference

The asymptotic result from the previous section can be used to motivate practical methods for large-sample inference using SDID. Under appropriate conditions, the estimator is asymptotically normal and zero centered; thus, if these conditions hold and we have a consistent estimator for its asymptotic variance V_{τ}, we can use conventional confidence intervals

$$\tau\in\hat{\tau}^{sdid}\pm z_{\alpha/2}\sqrt{\hat{V}_{\tau}} \tag{29}$$

to conduct asymptotically valid inference. In this section, we discuss three approaches to variance estimation for use in confidence intervals of this type.

The first proposal we consider, described in detail in Algorithm 2, involves a clustered bootstrap (Efron 1979) where we independently resample units. As argued in Bertrand, Duflo, and Mullainathan (2004), unit-level bootstrapping presents a natural approach to inference with panel data when repeated observations of the same unit may be correlated with each other. The bootstrap is simple to implement and, in our experiments, appears to yield robust performance in large panels. The main downside of the bootstrap is that it may be computationally costly as it involves running the full SDID algorithm for each bootstrap replication, and for large datasets this can be prohibitively expensive.

To address this issue we next consider an approach to inference that is more closely tailored to the SDID method and only involves running the full SDID algorithm once, thus dramatically decreasing the computational burden. Given weights $\hat{\omega}$ and $\hat{\lambda}$ used to get the SDID point estimate, Algorithm 3 applies the jackknife (Miller

ALGORITHM 2—BOOTSTRAP VARIANCE ESTIMATION

Data: $\mathbf{Y}, \mathbf{W}, B$
Result: Variance estimator $\hat{V}_\tau^{cb}$

for $i \leftarrow 1$ to B do
  Construct a bootstrap dataset $(\mathbf{Y}^{(b)}, \mathbf{W}^{(b)})$ by sampling N rows of
  $(\mathbf{Y}, \mathbf{W})$ with replacement.
  if the bootstrap sample has no treated units or no control units then
    Discard and resample (go to 2)
  end
  Compute the SDID estimator $\tau^{(b)}$ based on $(\mathbf{Y}^{(b)}, \mathbf{W}^{(b)})$
end
Define $\hat{V}_\tau^b = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{(b)})^2$;

ALGORITHM 3—JACKKNIFE VARIANCE ESTIMATION

Data: $\hat{\omega}$, $\hat{\lambda}$, $\mathbf{Y}$, $\mathbf{W}$, $\hat{\tau}$
Result: Variance estimator $\hat{V}_\tau$

for $i \leftarrow 1$ to N do
  Compute $\hat{\tau}^{(-i)}$: $\arg\min_{\tau,\{\alpha_j,\beta_j\}_{j\neq i,t}} \sum_{j\neq i,t}(\mathbf{Y}_{jt} - \alpha_j - \beta_t - \tau\mathbf{W}_{it})^2 \hat{\omega}_j \hat{\lambda}_t$
end
Compute $\hat{V}_\tau^{jack} = (N-1)N^{-1}\sum_{i=1}^{N}(\hat{\tau}^{(-i)} - \hat{\tau})^2$;

1974) to the weighted SDID regression (1), with the weights treated as fixed. The validity of this procedure is not implied directly by asymptotic linearity as in (27); however, as shown below, we still recover conservative confidence intervals under considerable generality.

THEOREM 2: *Suppose that the elements of* $\mathbf{L}$ *are bounded. Then, under the conditions of Theorem 1, the jackknife variance estimator described in Algorithm 3 yields conservative confidence intervals, i.e., for any* $0 < \alpha < 1$,

$$\liminf \Pr\left[\tau \in \hat{\tau}^{sdid} \pm z_{\alpha/2}\sqrt{\hat{V}_\tau^{jack}}\right] \geq 1 - \alpha. \tag{30}$$

Moreover, if the treatment effects $\boldsymbol{\tau}_{it} = \tau$ *are constant*[11] *and*

$$T_{post}N_{tr}^{-1}\|\hat{\lambda}_0 + \mathbf{L}_{tr,pre}\hat{\lambda}_{pre} - \mathbf{L}_{tr,post}\hat{\lambda}_{post}\|_2^2 \to_p 0, \tag{31}$$

that is, the time weights $\hat{\lambda}$ *are predictive enough on the exposed units, then the jackknife yields exact confidence intervals and* (30) *holds with equality.*

[11] When treatment effects are heterogeneous, the jackknife implicitly treats the estimand (15) as random whereas we treat it as fixed, thus resulting in excess estimated variance; see Imbens (2004) for further discussion.

ALGORITHM 4—PLACEBO VARIANCE ESTIMATION

Data: $\mathbf{Y}_{co,\cdot}, N_{tr}, B$
Result: Variance estimator $\hat{V}_\tau^{placebo}$

for $b \leftarrow 1$ to B do
  Sample N_{tr} out of the N_{co} control units without replacement to 'receive the placebo';
  Construct a placebo treatment matrix $\mathbf{W}_{co,\cdot}^{(b)}$ for the controls;
  Compute the SDID estimator $\hat{\tau}^{(b)}$ based on $(\mathbf{Y}_{co,\cdot}, \mathbf{W}_{co,\cdot}^{(b)})$;
end
Define $\hat{V}_\tau^{placebo} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{(b)})^2$;

In other words, we find that the jackknife is in general conservative and is exact when treated and control units are similar enough that time weights that fit the control units generalize to the treated units. This result depends on specific structure of the SDID estimator, and does not hold for related methods such as the SC estimator. In particular, an analogue to Algorithm 3 for SC would be severely biased upwards, and would not be exact even in the well-specified fixed effects model. Thus, we do not recommend (or report results for) this type of jackknifing with the SC estimator. We do report results for jackknifing DID since, in this case, there are no random weights $\hat{\omega}$ or $\hat{\lambda}$ and so our jackknife just amounts to the regular jackknife.

Now, both the bootstrap- and jackknife-based methods discussed so far are designed with the setting of Theorem 1 in mind, i.e., for large panels with many treated units. These methods may be less reliable when the number of treated units N_{tr} is small, and the jackknife is not even defined when $N_{tr} = 1$. However, many applications of SCs have $N_{tr} = 1$, e.g., the California smoking application from Section I. To this end, we consider a third variance estimator that is motivated by placebo evaluations as often considered in the literature on SCs (Abadie, Diamond, and Hainmueller 2010,;2015), and that can be applied with $N_{tr} = 1$. The main idea of such placebo evaluations is to consider the behavior of SC estimation when we replace the unit that was exposed to the treatment with different units that were not exposed.[12] Algorithm 4 builds on this idea, and uses placebo predictions using only the unexposed units to estimate the noise level, and then uses it to get $\hat{V}_\tau$ and build confidence intervals as in (29). See Bottmer et al. (2021) for a discussion of the properties of such placebo variance estimators in small samples.

Validity of the placebo approach relies fundamentally on homoskedasticity across units, because if the exposed and unexposed units have different noise distributions then there is no way we can learn V_τ from unexposed units alone. We also note that nonparametric variance estimation for treatment effect estimators is in general impossible if we only have one treated unit, and so homoskedasticity across units is effectively a necessary assumption in order for inference to be possible

[12] Such a placebo test is closely connected to permutation tests in randomization inference; however, in many SC applications, the exposed unit was not chosen at random, in which case placebo tests do not have the formal properties of randomization tests (Firpo and Possebom 2018, Hahn and Shi 2016), and so may need to be interpreted via a more qualitative lens.

TABLE 4

	Bootstrap			Jackknife			Placebo		
	SDID	SC	DID	SDID	SC	DID	SDID	SC	DID
1. Baseline	0.96	0.93	0.89	0.93	—	0.92	0.95	0.88	0.96
2. Gun law	0.97	0.96	0.92	0.94	—	0.93	0.94	0.95	0.93
3. Abortion	0.96	0.94	0.93	0.93	—	0.95	0.97	0.91	0.96
4. Random	0.96	0.96	0.92	0.93	—	0.94	0.96	0.96	0.94
5. Hours	0.92	0.96	0.94	0.89	—	0.95	0.91	0.90	0.96
6. Urate	0.78	0.74	0.38	0.71	—	0.42	0.74	0.77	0.41
7. $T_{post} = 1$	0.93	0.94	0.84	0.92	—	0.88	0.92	0.90	0.92
8. $N_{tr} = 1$	—	—	—	—	—	—	0.97	0.95	0.96
9. $T_{post} = N_{tr} = 1$	—	—	—	—	—	—	0.96	0.94	0.94
10. Resample, $N = 200$	0.94	0.96	0.92	0.95	—	0.93	0.96	0.95	0.94
11. Resample, $N = 400$	0.95	0.91	0.96	0.96	—	0.95	0.96	0.90	0.96
12. Democracy	0.93	0.96	0.55	0.94	—	0.59	0.98	0.97	0.79
13. Education	0.95	0.95	0.30	0.95	—	0.34	0.99	0.90	0.94
14. Random	0.93	0.95	0.89	0.96	—	0.91	0.95	0.94	0.91

Notes: Coverage results for nominal 95 percent confidence intervals in the CPS and Penn World Table simulation setting from Tables 2 and 3. The first three columns show coverage of confidence intervals obtained via the clustered bootstrap. The second set of columns show coverage from the jackknife method. The last set of columns show coverage from the placebo method. Unless otherwise specified, all settings have $N = 50$ and $T = 40$ cells, of which at most $N_{tr} = 10$ units and $T_{post} = 10$ periods are treated. In rows 7–9, we reduce the number of treated cells. In rows 10 and 11, we artificially make the panel larger by adding rows, which makes the assumption that the number of treated units is small relative to the number of control units more accurate. (We set N_{tr} to 10 percents of the total number of units.) We do not report jackknife and bootstrap coverage rates for $N_{tr} = 1$ because the estimators are not well-defined. We do not report jackknife coverage rates for SC because, as discussed in the text, the variance estimator is not well justified in this case. All results are based on 400 simulation replications.

here.[13] Algorithm 4 can also be seen as an adaptation of the method of Conley and Taber (2011) for inference in DID models with few treated units and assuming homoskedasticity, in that both rely on the empirical distribution of residuals for placebo-estimators run on control units to conduct inference. We refer to Conley and Taber (2011) for a detailed analysis of this class of algorithms.

Table 4 shows the coverage rates for the experiments described in Section IIA and IIB, using Gaussian confidence intervals (29) with variance estimates obtained as described above. In the case of the SDID estimation, the bootstrap estimator performs particularly well, yielding nearly nominal 95 percent coverage, while both placebo and jackknife variance estimates also deliver results that are close to the nominal 95 percent level. This is encouraging, and aligned with our previous observation that the SDID estimator appeared to have low bias. That being said, when assessing the performance of the placebo estimator, recall that the data in Section IIA was generated with noise that is both Gaussian and homoskedastic across units—which were assumptions that are both heavily used by the placebo estimator.

In contrast, we see that coverage rates for DID and SC can be relatively low, especially in cases with significant bias such as the setting with the state unemployment rate as the outcome. This is again in line with what one may have expected based on the distribution of the errors of each estimator as discussed in Section IIA, e.g., in

[13] In Theorem 1, we also assumed homoskedasticity. In contrast to the case of placebo inference, however, it's likely that a similar result would also hold without homoskedasticity; homoskedasticity is used in the proof essentially only to simplify notation and allow the use of concentration inequalities which have been proven in the homoskedastic case but can be generalized.

Figure 2: If the point estimates $\hat{\tau}$ from DID and SC are dominated by bias, then we should not expect confidence intervals that only focus on variance to achieve coverage.

V. Related Work

Methodologically, our work draws most directly from the literature on SC methods, including Abadie and Gardeazabal (2003); Abadie, Diamond, and Hainmueller (2010, 2015); Abadie and L'Hour (2016); Doudchenko and Imbens (2016); and Ben-Michael, Feller, and Rothstein (2018). Most methods in this line of work can be thought of as focusing on constructing unit weights that create comparable (balanced) treated and control units, without relying on any modeling or weighting across time. Ben-Michael, Feller, and Rothstein (2018) is an interesting exception. Their augmented SC estimator, motivated by the augmented inverse-propensity weighted estimator of Robins, Rotnitzky, and Zhao (1994), combines SC weights with a regression adjustment for improved accuracy. (See also Kellogg et al. 2020 which explicitly connects SC to matching). They focus on the case of $N_{tr} = 1$ exposed units and $T_{post} = 1$ postexposure periods, and their method involves fitting a model for the conditional expectation $m(\cdot)$ for Y_{iT} in terms of the lagged outcomes $\mathbf{Y}_{i,pre}$, and then using this fitted model to "augment" the basic SC estimator as follows:

$$\hat{\tau}_{asc} = Y_{NT} - \left(\sum_{i=1}^{N-1} \hat{\omega}_i^{sc} Y_{iT} + \left(\hat{m}\big(\mathbf{Y}_{N,pre}\big) - \sum_{i=1}^{N-1} \hat{\omega}_i^{sc} \hat{m}\big(\mathbf{Y}_{i,pre}\big) \right) \right). \tag{32}$$

Despite their different motivations, the augmented SC and SDID methods share an interesting connection: with a linear model $m(\cdot)$, $\hat{\tau}_{sdid}$ and $\hat{\tau}_{asc}$ are very similar. In fact, had we fit $\hat{\omega}^{sdid}$ without intercept, they would be equivalent for $\hat{m}(\cdot)$ fit by least squares on the controls, imposing the constraint that its coefficients are nonnegative and to sum to one, that is, for $\hat{m}(\mathbf{Y}_{i,pre}) = \hat{\lambda}_0^{sdid} + \mathbf{Y}_{i,pre} \hat{\lambda}_{pre}^{sdid}$. This connection suggests that weighted two-way bias-removal methods are a natural way of working with panels where we want to move beyond simple DID approaches.

We also note recent work of Roth (2018) and Rambachan and Roth (2019), who focus on valid inference in DID settings when users look at past outcomes to check for parallel trends. Our approach uses past data not only to check whether the trends are parallel, but also to construct the weights to make them parallel. In this setting, we show that one can still conduct valid inference, as long as N and T are large enough and the size of the treatment block is small.

In terms of our formal results, our paper fits broadly in the literature on panel models with interactive fixed effects and the matrix completion literature (Athey et al. 2021; Bai 2009; Moon and Weidner 2015, 2017; Robins 1985; Xu 2017). Different types of problems of this form have a long tradition in the econometrics literature, with early results going back to Ahn, Lee, and Schmidt (2001); Chamberlain (1992); and Holtz-Eakin, Newey, and Rosen (1988) in the case of finite-horizon panels (i.e., in our notation, under asymptotics where T is fixed and only $N \to \infty$). More recently, Freyberger (2018) extended the work of Chamberlain (1992) to a setting that's closely related to ours, and emphasized the role of the past outcomes for constructing moment restrictions in the fixed-T setting.

Freyberger (2018) attains identification by assuming that the errors $\mathbf{E}_{it}$ are uncorrelated, and thus past outcomes act as valid instruments. In contrast, we allow for correlated errors within rows, and thus need to work in a large-T setting.

Recently, there has considerable interest in models of type (10) under asymptotics where both N and T get large. One popular approach, studied by Bai (2009) and Moon and Weidner (2015, 2017), involves fitting (10) by "least squares," i.e., by minimizing squared-error loss while constraining $\hat{\mathbf{L}}$ to have bounded rank R. While these results do allow valid inference for τ, they require strong assumptions. First, they require the rank of $\mathbf{L}$ to be known a priori (or, in the case of Moon and Weidner 2015, require a known upper bound for its rank), and second, they require a $\beta_{\min}$-type condition whereby the normalized nonzero singular values of $\mathbf{L}$ are well separated from zero. In contrast, our results require no explicit limit on the rank of $\mathbf{L}$ and allow for $\mathbf{L}$ to have to have positive singular values that are arbitrarily close to zero, thus suggesting that the SDID method may be more robust than the least squares method in cases where the analyst wishes to be as agnostic as possible regarding properties of $\mathbf{L}$.[14]

Athey et al. (2021); Amjad, Shah, and Shen (2018); Moon and Weidner (2018;, and Xu (2017) build on this line of work, and replace the fixed-rank constraint with data-driven regularization on $\hat{\mathbf{L}}$. This innovation is very helpful from a computational perspective; however, results for inference about τ that go beyond what was available for least squares estimators are currently not available. We also note recent papers that draw from these ideas in connection to SC type analyses, including Chan and Kwok (2020) and Gobillon and Magnac (2016). Finally, in a paper contemporaneous to ours, Agarwal et al. (2019) provide improved bounds from principal component regression in an errors-in-variables model closely related to our setting, and discuss implications for estimation in SC type problems. Relative to our results, however, Agarwal et al. (2019) still require assumptions on the behavior of the small singular values of $\mathbf{L}$, and do not provide methods for inference about τ.

In another direction, several authors have recently proposed various methods that implicitly control for the systematic component $\mathbf{L}$ in models of time (10). In one early example, Hsiao, Ching, and Ki Wan (2012) start with a factor model similar to ours and show that under certain assumptions it implies the moment condition

$$Y_{Nt} = a + \sum_{j=1}^{N-1} \beta_j Y_{jt} + \epsilon_{Nt}, \qquad E\left[\varepsilon_{Nt} \middle| \left\{Y_{jt}\right\}_{j=1}^{N-1}\right] = 0, \tag{33}$$

for all $t = 1, \ldots, T$. The authors then estimate β_j by (weighted) ordinary least squares. This approach is further refined by Li and Bell (2017), who additionally propose to penalizing the coefficients β_j using the lasso (Tibshirani 1996). In a recent paper, Chernozhukov, Wüthrich, and Zhu (2018) use the model (33) as a starting point for inference.

While this line of work shares a conceptual connection with us, the formal setting is very different. In order to derive a representation of the type (33), one essentially needs to assume a random specification for (10) where both $\mathbf{L}$ and $\mathbf{E}$ are stationary

[14] By analogy, we also note that, in the literature on high-dimensional inference, methods that do no assume a uniform lower bound on the strength of nonzero coefficients of the signal vector are generally considered more robust than ones that do (e.g., Belloni, Chernozhukov, and Hansen 2014; Zhang and Zhang 2014).

in time. Li and Bell (2017) explicitly assumes that the outcomes $\mathbf{Y}$ themselves are weakly stationary, while Chernozhukov, Wüthrich, and Zhu (2018) makes the same assumption to derive the results that are valid under general misspecification. In our results, we do not assume stationarity anywhere: $\mathbf{L}$ is taken as deterministic and the errors $\mathbf{E}$ may be nonstationary. Moreover, in the case of most SC and DID analyses, we believe stationarity to be a fairly restrictive assumption. In particular, in our model, stationarity would imply that a simple pre-post comparison for exposed units would be an unbiased estimator of τ and, as a result, the only purpose of the unexposed units would be to help improve efficiency. In contrast, in our analysis, using unexposed units for double differencing is crucial for identification.

Ferman and Pinto (2019) analyze the performance of SC estimator using essentially the same model as we do. They focus on the situations where N is small, while T_{pre} (the number of control periods) is growing. They show that unless time factors have strong trends (e.g., polynomial) the SC estimator is asymptotically biased. Importantly Ferman and Pinto (2019) focus on the standard SC estimator, without time weights and regularization, but with an intercept in the construction of the weights.

Finally, from a statistical perspective, our approach bears some similarity to the work on "balancing" methods for program evaluation under unconfoundedness, including Athey, Imbens, and Wager (2018); Graham, Pinto, and Egel (2012); Hirshberg and Wager (2017); Imai and Ratkovic (2014); Kallus (2020); Zhao (2019); and Zubizarreta (2015). One major result of this line of work is that, by algorithmically finding weights that balance observed covariates across treated and control observations, we can derive robust estimators with good asymptotic properties (such as efficiency). In contrast to this line of work, rather than balancing observed covariates, we here need to balance unobserved factors $\mathbf{\Gamma}$ and $\mathbf{\Upsilon}$ in (10) to achieve consistency; and accounting for this forces us to follow a different formal approach than existing studies using balancing methods.

Appendix. Staggered Adoption

In the paper so far we have focused on the case where some units start receiving the treatment at a common point in time, what Athey et al. (2021) call block assignment. Under block assignment the $N \times T$ matrix of treatment assignments $\mathbf{W}$ has the form like the following matrix, where units 3–6 all adopt the treatment in period 5:

$$\mathbf{W} = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

This is a common setting, but there are other settings that are of interest. Another important special case is that of *staggered adoption* (e.g.,Athey and Imbens 2021) with multiple dates at which the treatment is started. For example, in the following

assignment matrix units 5 and 6 adopt the treatment in period 3, and units 3 and 4 adopt the treatment in period 5 (and units 1 and 2 never adopt the treatment):

$$\mathbf{W} = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 5 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

With staggered adoption the weighted DID regression approach in SDID does not work directly. However, there are various alternatives. Here we discuss a simple modification to estimate the average treatment effect for the treated in that setting by applying the SDID estimator repeatedly, once for every adoption date. An alternative is the procedure developed in Ben-Michael, Feller, and Rothstein (2019). In the example above with two adoption dates, we can create two assignment matrices, $\mathbf{W}^1$ and $\mathbf{W}^2$, that both fit into the block assignment setting. We can then apply the SDID estimator to both samples, and calculate a weighted average of the two estimators, with the weight equal to the fraction of treated unit/time-period pairs in each of the two samples. In the above example, the first sample would consist of units 1, 2, 5 and 6, and the second sample would consist of units 1, 2, 3, and 4, as illustrated in the two assignment matrices below:

$$\mathbf{W}^1 = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{W}^2 = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Alternatively we can create the two samples by splitting the data up by time periods. In that case the first sample would consist of time periods 1, 2, 3, and 4, and the second sample would consist of time periods 1, 2, 5, 6, and 7, as illustrated below:

$$\mathbf{W}^1 = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 1 & 1 \\ 6 & 0 & 0 & 1 & 1 \end{pmatrix}, \qquad \mathbf{W}^2 = \begin{pmatrix} & 1 & 2 & 5 & 6 & 7 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 1 & 1 & 1 \\ 5 & 0 & 0 & 1 & 1 & 1 \\ 6 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

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.

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

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 113–32.

Abadie, Alberto, and Jérémy L'Hour. 2016. "A Penalized Synthetic Control Estimator for Disaggregated Data." Unpublished.

Agarwal, Anish, Devavrat Shah, Dennis Shen, and Dogyoon Song. 2019. "On Robustness of Principal Component Regression." arXiv preprint arXiv:1902.10920.

Ahn, Seung Chan, Young Hoon Lee, and Peter Schmidt. 2001. "GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects." *Journal of Econometrics* 101 (2): 219–55.

Amjad, Muhammad, Devavrat Shah, and Dennis Shen. 2018. "Robust Synthetic Control." *Journal of Machine Learning Research* 19 (22): 1–51.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Replication Data for: Synthetic Difference-in-Differences." American Economic Association [publisher], Inter-university Consortium for Social and Political Research [distributor]. https://doi.org/10.3886/E146381V1.

Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67 (4): 648–60.

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. https://doi.org/10.1080/01621459.2021.1891924.

Athey, Susan, and Guido W. Imbens. 2021. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.10.012.

Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 80 (4): 597–623.

Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77 (4): 1229–79.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár. 2012. "Clustering, Spatial Correlations, and Randomization Inference." *Journal of the American Statistical Association* 107 (498): 578–91.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81 (2): 608–50.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2018. "The Augmented Synthetic Control Method." arXiv preprint arXiv:1811.04170v1.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2019. "Synthetic Controls and Weighted Event Studies with Staggered Adoption." arXiv preprint arXiv:1912.03290v1.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–75.

Borusyak, Kirill, and Xavier Jaravel. 2016. "Revisiting Event Study Designs." Unpublished.

Bottmer, Lea, Guido Imbens, Jann Spiess, and Merrill Warnick. 2021. "A Design-Based Perspective on Synthetic Control Methods." arXiv preprint arXiv:2101.09398v1.

Callaway, Brantly, and Pedro H. C. Sant'anna. 2020. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.12.001.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relation* 43 (2): 245–57.

Chamberlain, Gary. 1992. "Efficiency Bounds for Semiparametric Regression." *Econometrica* 60 (3): 567–96.

Chan, Mark K., and Simon Kwok. 2020. "The PCDID Approach: Difference-in-Differences When Trends Are Potentially Unparallel and Stochastic." Unpublished.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. 2018. "Inference on Average Treatment Effects in Aggregate Panel Data Settings." arXiv preprint arXiv:1812.10820v1.
Conley, Timothy G., and Christopher R. Taber. 2011. "Inference with 'Difference in Difference' with a Small Number of Policy Changes." *Review of Economics and Statistics* 93 (1): 113–25.
Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods." *AEA Papers and Proceedings* 110: 42–48.
de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.
Doudchenko, Nikolay, and Guido W. Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." NBER Working Paper 22791.
Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7 (1): 1–26.
Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." *American Economic Review* 105 (10): 3150–82. Available for download at www.ggdc.net/pwt.
Ferman, Bruno, and Cristine Pinto. 2019. "Synthetic Controls with Imperfect Pre-treatment Fit." arXiv preprint arXiv:1911.08521v1.
Firpo, Sergio, and Vitor Possebom. 2018. "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets." *Journal of Causal Inference* 6 (2): Article 20160026.
Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. "Pre-event Trends in the Panel Event-Study Design." *American Economic Review* 109 (9): 3307–38.
Freyberger, Joachim. 2018. "Non-parametric Panel Data Models with Interactive Fixed Effects." *Review of Economic Studies* 85 (3): 1824–51.
Gobillon, Laurent, and Thierry Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98 (3): 535–51.
Graham, Bryan S., Cristine Campos De Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *Review of Economic Studies* 79 (3): 1053–79.
Hahn, Jinyong, and Ruoyao Shi. 2016. "Synthetic Control and Inference." Unpublished.
Hirshberg, David A. 2021. "Least Squares with Error in Variables." arXiv preprint arXiv:2104.08931v1.
Hirshberg, David A., and Stefan Wager. 2017. "Augmented Minimax Linear Estimation." arXiv preprint arXiv:1712.00038.
Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56 (6): 1371–95.
Hsiao, Cheng, H. Steve Ching, and Shui Ki Wan. 2012. "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China." *Journal of Applied Econometrics* 27 (5): 705–40.
Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 76 (1): 243–63.
Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86 (1): 4–29.
Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.
Kallus, Nathan. 2020. "Generalized Optimal Matching Methods for Causal Inference." *Journal of Machine Learning Research* 21 (62): 1–54.
Kellogg, Maxwell, Magne Mogstad, Guillaume Pouliot, and Alexander Torgovitsky. 2020. "Combining Matching and Synthetic Control to Trade Off Biases from Extrapolation and Interpolation." NBER Working Paper 26624.
Li, Kathleen T., and David R. Bell. 2017. "Estimation of Average Treatment Effects with Panel Data: Asymptotic Theory and Implementation." *Journal of Econometrics* 197 (1): 65–75.
Miller, Rupert G. 1974. "The Jackknife - A Review." *Biometrika* 61 (1): 1–15.
Moon, Hyungsik Roger, and Martin Weidner. 2017. "Dynamic Linear Panel Regression Models with Interactive Fixed Effects." *Econometric Theory* 33 (1): 158–95.
Moon, Hyungsik Roger, and Martin Weidner. 2015. "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects." *Econometrica* 83 (4): 1543–79.
Moon, Hyungsik Roger, and Martin Weidner. 2018. "Nuclear Norm Regularized Estimation of Panel Regression Models." arXiv preprint arXiv:1810.10987v1.
National Bureau of Economic Research. 2021. "Merged Outgoing Rotation Groups (MORG)." https://data.nber.org/morg/annual (accessed August 1, 2021).
Orzechowski, & Walker. 2005. *The Tax Burden on Tobacco.* Historical Compilation, Vol. 40, Arlington, VA: Orzechowski & Walker.

Peri, Giovanni, and Vasil Yasenov. 2019. "The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift." *Journal of Human Resources* 54 (2): 267–309.

Rambachan, Ashesh, and Jonathan Roth. 2019. "An Honest Approach to Parallel Trends." Unpublished.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Robins, Philip K. 1985. "A Comparison of the Labor Supply Findings from the Four Negative Income Tax Experiments." *Journal of Human Resources* 20 (4): 567–82.

Roth, Jonathan. 2018. "Pre-test with Caution: Event-Study Estimates after Testing for Parallel Trends." Unpublished.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association* 94 (448): 1096–1120.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B* (*Methodological*) 58 (1): 267–88.

Vershynin, Roman. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge, UK: Cambridge University Press.

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

Zhang, Cun-Hui, and Stephanie S. Zhang. 2014. "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models." *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) 76 (1): 217–42.

Zhao, Qingyuan. 2019. "Covariate Balancing Propensity Score by Tailored Loss Functions." *Annals of Statistics* 47 (2): 965–93.

Zubizarreta, José R. 2015. "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110 (511): 910–22.