Efficient Estimation for Staggered Rollout Designs

Jonathan Roth
Brown University

Pedro H. C. Sant’Anna
Emory University

We study estimation of causal effects in staggered-rollout designs—that is, settings where there is staggered treatment adoption and the timing of treatment is as good as randomly assigned. We derive the most efficient estimator in a class of estimators that nests several popular generalized difference-in-differences methods. A feasible plug-in version of the efficient estimator is asymptotically unbiased, with efficiency (weakly) dominating that of existing approaches. We provide both $t$-based and permutation-test-based methods for inference. In an application to a training program for police officers, confidence intervals for the proposed estimator are as much as eight times shorter than those for existing approaches.

I. Introduction

Researchers are often interested in the causal effect of a treatment that is first implemented for different units at different times. Staggered rollouts

We are grateful to Bocar Ba, Yuehao Bai, Brantly Callaway, Ivan Canay, Clément de Chaisemartin, Jen Doleac, Peng Ding, Avi Feller, Ryan Hill, Lihua Lei, David McKenzie, Emily Owens, Ashesh Rambachan, Roman Rivera, Evan Rose, Adrienne Sabyet, Jesse Shapiro, Yotam Shem-Tov, Dylan Small, Ariella Kahn-Lang Spitzer, Sophie Sun, and seminar participants at Columbia, the European Union Joint Research Centre, Insper Institute of Education and Research, Notre Dame, the Paris School of Economics/Center de Recherche

Electronic publication October 16, 2023
Journal of Political Economy Microeconomics, volume 1, number 4, November 2023.
© 2023 The University of Chicago. All rights reserved. Published by The University of Chicago Press.
https://doi.org/10.1086/720581

669
are frequently analyzed with methods that extend the simple two-period difference-in-differences (DiD) estimator to the staggered setting, such as two-way fixed-effects (TWFE) regression estimators and recently proposed alternatives that yield more intuitive causal parameters under treatment effect heterogeneity (de Chaisemartin and D’Haultfoeuille 2020; Callaway and Sant’Anna 2021; Sun and Abraham 2021). The validity of these estimators depends on a parallel-trends assumption.

However, researchers often justify this assumption by arguing that the timing of the treatment is as good as randomly assigned. In some settings, such as our application to the rollout of a training program for police officers, the timing of the treatment is explicitly randomized. In other settings, treatment timing is not explicitly randomized, but the researcher argues that it is due to idiosyncratic quasi-random factors. For example, Deshpande and Li (2019, 226) justify the use of a DiD design comparing areas whose social security office closed at different times by arguing that the “timing of the closings appears to be effectively random,” as evidenced by the fact that observable characteristics are balanced across places with closures at different times. DiD and related methods have also been used to exploit the quasi-random timing of parental deaths (Nekoei and Seim 2023), health shocks (Fadlon and Nielsen 2021), and stimulus payments (Parker et al. 2013), among others.

In this paper, we show that if treatment timing is as good as randomly assigned, one can obtain estimates more precise than those provided by DiD-based methods. We derive the most efficient estimator in a large class of estimators that nests many existing DiD-based approaches, and we show how to conduct both \(t\)-based and permutation-based inference. In settings where treatment timing is as good as random, our efficient estimator has the scope to substantially reduce standard errors, as illustrated in our simulations and application below.

We begin by introducing a design-based framework that formalizes the notion that treatment timing is (quasi-)randomly assigned. There are \(T\) periods, and unit \(i\) is first treated in period \(G_i \in G \subseteq \{1, \ldots, T, \infty\}\), with \(G_i = \infty\) denoting that \(i\) is never treated (or treated after period \(T\)). We make

---

1 When treatment is as good as randomly assigned, other methods (e.g., simple comparisons of means) are available to estimate average treatment effects. DiD-based methods have nevertheless been recommended for randomized rollouts to improve efficiency (Xiong et al. 2019) and to transparently aggregate treatment effect heterogeneity (Lindner and McConnell 2021).
two key assumptions in this model. First, we assume that the treatment timing $G_i$ is (quasi-)randomly assigned, in the sense that any permutation of the observed vector of treatment start dates is equally likely to occur. Second, we rule out anticipatory effects of treatment—for example, a unit’s outcome in period 2 does not depend on whether it was first treated in period 3 or in period 4.

Within this framework, we show that pretreatment outcomes play a role similar to that of fixed covariates in a randomized experiment and that generalized DiD estimators can be viewed as applying a crude form of covariate adjustment. To develop the intuition, it is instructive to first consider the special case where we observe data for two periods ($T = 2$), some units are first treated in period 2 ($G_i = 2$), and the remaining units are treated in a later period or never treated ($G_i = \infty$). This special case is analogous to conducting a randomized experiment in period 2, with the outcome in period 1 serving as a pretreatment covariate. The DiD estimator is $\hat{\theta}_{DiD}^{\ast} = (\bar{Y}_{22} - \bar{Y}_{\infty}) - (\bar{Y}_{12} - \bar{Y}_{\infty})$, where $\bar{Y}_{gt}$ is the mean outcome for treatment group $g$ at period $t$. It is clear that $\hat{\theta}_{DiD}^{\ast}$ is a special case of the class of estimators

$$\hat{\theta}_{g} = \left(\bar{Y}_{22} - \bar{Y}_{2g}\right) - \beta \left(\bar{Y}_{12} - \bar{Y}_{1g}\right)$$

that adjust the posttreatment difference in means by $\beta$ times the pretreatment difference in means. Under the assumption of (quasi-)random treatment timing, $\hat{\theta}_{g}$ is unbiased for the average treatment effect (ATE) for any $\beta$, since the posttreatment difference in means is unbiased for the ATE and the pretreatment difference in means is mean zero. The value of $\beta$ that minimizes the variance of the estimator depends on the covariances of the potential outcomes between periods, however. Intuitively, we want to put more weight on lagged outcomes when they are more informative about posttreatment outcomes. DiD, which imposes the fixed weight $\beta = 1$, will thus generally be inefficient, and one can obtain an (asymptotically) more efficient estimator by estimating the optimal weights from the data. In this special two-period case, the form of the efficient estimator follows from Lin (2013), who studied efficient covariate adjustment in cross-sectional randomized experiments; see also McKenzie (2012), who noted that the two-period DiD estimator may be inefficient in experiments.

Our main theoretical results extend this logic to the case of staggered treatment timing, providing formal methods for estimation and inference. We begin by introducing a flexible class of causal parameters that can highlight treatment effect heterogeneity across both calendar time and time since treatment. Following Athey and Imbens (2022), we define $\tau_{t;g}^{g'}$ to be the average effect on the outcome in period $t$ of changing the initial treatment date from $g'$ to $g$. For example, in the simple two-period case described above, $\tau_{2;2x}$ corresponds with the ATE on the second-period staggered rollout designs.
outcome of being treated in period 2 relative to never being treated. We then consider the class of estimands that are linear combinations of these building blocks, \( \theta = \sum_{i \in d} a_{i \in g} \tau_{i \in g} \). Our framework thus allows for arbitrary treatment effect dynamics and accommodates a variety of ways of summarizing these dynamic effects, including several aggregation schemes proposed in the recent literature.

We then consider the large class of estimators that start with a sample analogue to the target parameter and adjust by a linear combination of differences in pretreatment outcomes. More precisely, we consider estimators of the form \( \hat{\theta}_b = \sum_{i \in d} a_{i \in g} \hat{\tau}_{i \in g} - \hat{X}'\beta \), where the first term is a sample analogue to \( \theta \) and the second term adjusts linearly using a vector \( \hat{X} \) that compares outcomes for cohorts treated at different dates at points in time before either was treated. For example, in the simple two-period case described above, \( \hat{X} = \bar{Y}_{12} - \bar{Y}_{1\infty} \) is the difference in means in period 1. We show that several estimators for the staggered setting are part of this class for an appropriately defined estimand and \( \hat{X} \), including the TWFE estimator as well as recent procedures proposed by de Chaisemartin and D’Haultfoeuille (2020), Callaway and Sant’Anna (2021), and Sun and Abraham (2021). All estimators of this form are unbiased for \( \theta \) under the assumptions of (quasi-)random treatment timing and no anticipation.

We then derive the most efficient estimator in this class. The optimal coefficient \( \hat{\beta}^* \) depends on covariances between the potential outcomes over time, and thus the estimators previously proposed in the literature will be efficient only for special covariance structures. Although the covariances of the potential outcomes are generally not known ex ante, one can estimate a “plug-in” version of the efficient estimator that replaces the “oracle” coefficient \( \hat{\beta}^* \) with a sample analogue \( \hat{\beta}^* \). We show that the plug-in efficient estimator is asymptotically unbiased and as efficient as the oracle estimator under large-population asymptotics similar to those in Lin (2013) and Li and Ding (2017) for covariate adjustment in cross-sectional experiments.

Our results suggest two complementary approaches to inference. First, we show that the plug-in efficient estimator is asymptotically normally distributed in large populations, which allows for asymptotically valid confidence intervals (CIs) of the familiar form \( \hat{\theta}_{b^*} \pm 1.96\hat{\sigma}_e \). Second, an appealing feature of our (quasi-)random treatment timing framework is that it permits us to construct Fisher randomization tests (FRTs), also known as permutation tests. Following Wu and Ding (2021) and Zhao and Ding (2021) for cross-sectional randomized experiments, we consider FRTs based on a studentized version of our efficient estimator. These FRTs have the dual advantages that they are finite-sample exact under the

\[ As is common in finite-population settings, the covariance estimate may be conservative if there are heterogeneous treatment effects. \]
sharp null of no treatment effects and asymptotically valid for the weak null of no average effects. In a Monte Carlo study calibrated to our application, we find that both the $t$-based and FRT-based approaches yield reliable inference and that CIs based on the plug-in efficient estimator are substantially shorter than those for the procedures of de Chaisemartin and D’Haultfoeuille (2020), Callaway and Sant’Anna (2021), and Sun and Abraham (2021). 3

As an illustration of our method and stand-alone empirical contribution, we revisit the randomized rollout of a procedural-justice training program for police officers in Chicago. The original study by Wood, Tyler, and Papachristos (2020) found large and statistically significant reductions in complaints and officer use of force, and these findings were influential in policy debates about policing (Doleac 2020). Unfortunately, an earlier version of our analysis revealed a statistical error in the analysis of Wood, Tyler, and Papachristos (2020) that led their estimates to be inflated. In Wood et al. (2020), we collaborated with the original authors to correct this error. Using the estimator of Callaway and Sant’Anna (2021), we found no significant effects on complaints against police officers and borderline significant effects on officer use of force, but with wide CIs that included both near-zero and meaningfully large treatment effects estimates. We find that the use of the methodology proposed in this paper allows us to obtain substantially more precise estimates of the effect of the training program. Although we again find no statistically significant effects on complaints and borderline significant effects on force, the standard errors from using our methodology are between 1.4 and 8.4 times smaller than those from the Callaway and Sant’Anna (2021) estimator used in Wood et al. (2020). For complaints, for example, we are able to rule out reductions larger than 13% of the pretreatment mean using our proposed estimator, compared with an upper bound of 33% in the previous analysis.

**Related literature.**—This paper contributes to an active literature on DiD and related methods in settings with staggered treatment timing. Several recent papers have demonstrated the failures of TWFE models to recover a sensible causal estimand under treatment effect heterogeneity and have proposed alternative estimators with better properties (de Chaisemartin and D’Haultfoeuille 2020; Callaway and Sant’Anna 2021; Goodman-Bacon 2021; Sun and Abraham 2021; Borusyak, Jaravel, and Spiess, forthcoming). Most of this literature has focused on obtaining consistent estimates under a generalized parallel-trends assumption, whereas we focus on efficient estimation under the stronger assumption of (quasi-)random treatment timing. Our proposed efficient estimator can help to improve precision relative to DiD methods in settings where the researcher believes

---

3 The staggered R and Stata packages allow for easy implementation of the plug-in efficient estimator; see https://github.com/jonathandroth/staggered and https://github.com/mcaceresh/stata-staggered, respectively.
that treatment timing is as good as randomly assigned, but, unlike other estimators in the literature, it will not be applicable in settings where the researcher is confident in parallel trends but not (quasi-)random treatment timing. For example, our random-treatment-timing assumption requires that pretreatment outcomes and fixed covariates should be balanced across groups treated at different times, whereas this is not strictly required by the parallel-trends assumption. See remark 2 for further discussion.

Two related papers that have studied (quasi-)random treatment timing are Athey and Imbens (2022) and Shaikh and Toulis (2021). The former studies a model of random treatment timing similar to ours but focuses on the interpretation of the TWFE estimand. The latter paper adopts a different framework of randomization, in which treatment timing is random only conditional on observables and no two units can be treated at the same time. Neither paper considers the efficient choice of estimator, as we do.

Our technical results extend results in statistics on efficient covariate adjustment in cross-sectional experiments (Freedman 2008a, 2008b; Lin 2013; Li and Ding 2017) to the setting of staggered treatment timing, where pretreatment outcomes play a role similar to that of fixed covariates in a cross-sectional experiment. In the special two-period case, our proposed estimator reduces to Lin’s (2013) efficient estimator, treating the lagged outcome as a fixed covariate. Our results are also related to McKenzie (2012), who showed that DiD is inefficient under random treatment assignment in a two-period model with homogeneous treatment effects; see remark 3 (sec. II.E.1) for additional details. We note that the notion of efficiency studied in this paper is efficiency in the class of estimators of the form given in equation (1), rather than semiparametric efficiency as in, for example, Hahn (1998) and Sant’Anna and Zhao (2020). We are not aware of a notion of semiparametric efficiency for design-based models such as ours but consider this an interesting topic for future work.

Our paper also relates to the literature on clinical trials using a stepped-wedge design, which is a randomized staggered rollout in which all units are ultimately treated (e.g., Brown and Lilford 2006). Until recently, this literature has focused on estimation using mixed-effects regression models. Lindner and McConnell (2021) point out, however, that such models may be difficult to interpret under treatment effect heterogeneity, and they recommend using DiD-based approaches like that of Sun and Abraham (2021) instead. Our approach has the potential to offer large gains in precision relative to such DiD-based approaches. Our paper is also complementary to Ji et al. (2017), who propose using randomization-based inference procedures to test Fisher’s sharp null hypothesis in stepped-wedge designs. By contrast, we consider Neymanian inference on ATEs and also show that an FRT with a studentized test statistic is both finite-sample exact for the sharp null and asymptotically valid for inference on average effects.
Finally, our work is related to Xiong et al. (2019) and Basse, Ding, and Toulis (2023), who consider the optimal design of a staggered-rollout experiment to maximize the efficiency of a fixed estimator. By contrast, we solve for the most efficient estimator, given a fixed experimental design.

II. Model and Theoretical Results

A. Model

There is a finite population of $N$ units. We observe data for $T$ periods, $t = 1, ..., T$. A unit’s treatment status is denoted by $G_i \in \mathcal{G} \subseteq \{1, ..., T, \infty\}$, where $G_i$ is the first period in which unit $i$ is treated and $G_i = \infty$ denotes that a unit is never treated (or treated after period $T$). Our framework accommodates but does not require there to be never-treated units—it could be that $\infty \notin \mathcal{G}$, in which case all units are eventually treated (a stepped-wedge design). We assume that treatment is an absorbing state. We denote by $Y_{it}(g)$ the potential outcome for unit $i$ in period $t$ when treatment starts at time $g$, and we define the vector $Y_i(g) = (Y_{i1}(g), ..., Y_{iT}(g))^\prime \in \mathbb{R}^T$. We let $D_{ig} = 1[G_i = g]$. The observed vector of outcomes for unit $i$ is then $Y_i = \Sigma_g D_{ig} Y_i(g)$.

Following Neyman (1990 [1923]) for randomized experiments and Athey and Imbens (2022) for settings with staggered treatment timing, our model is design based: we treat as fixed (or condition on) the potential outcomes and the number of units first treated at each period ($N_g$). The only source of uncertainty in our model comes from the vector of times at which units are first treated, $G = (G_1, ..., G_N)^\prime$, which is stochastic.

Remark 1 (Design-based uncertainty). Design-based models are particularly attractive in settings where it is difficult to define the superpopulation, such as when all 50 states are observed (Manski and Pepper 2018) or, in our application, where the near-universe of police officers in Chicago is observed. Even when there is a superpopulation, the design-based view allows for valid inference on the sample ATE; see Abadie et al. (2020), Sekhon and Shem-Tov (2021) for additional discussion.

Our first main assumption is that the treatment timing is (quasi-)randomly assigned, meaning that any permutation of the treatment timing vector is equally likely.

Assumption 1 (Random treatment timing). Let $D$ be the random $N \times |\mathcal{G}|$ matrix with $(i, g)$-th element $D_{ig}$. Then, $P(D = d) = (\prod_{g \in \mathcal{G}} N_g !)/N!$ if $\sum_{g \in \mathcal{G}} d_g = N_g$ for all $g$, and zero otherwise.

\[ \text{If treatment turns on and off, the parameters we estimate can be viewed as the intent-to-treat effect of first being treated at a particular date; see de Chaisemartin and D’Haultfoeuille (2021) and Sun and Abraham (2021) for related discussion for DiD models.} \]
We note that assumption 1 will hold by design in settings where the researcher randomly assigns individuals to treatment start dates. It can also hold in quasi-experimental contexts if the idiosyncratic factors that determine treatment timing render any permutation of the treatment start dates to be equally likely; see Borusyak and Hull (2020) and Rambachan and Roth (2020) for additional discussion of “quasi-random” treatment assignment. We discuss extensions to clustered and conditional random assignment of treatment timing in section II.H.

Remark. 2 (Comparison to parallel-trends assumption). Technically speaking, the random-timing assumption in assumption 1 is stronger than the usual parallel-trends assumption, which requires only that treatment probabilities are orthogonal to trends in the potential outcomes. Assumption 1 thus may not be plausible in all settings where researchers use DiD methods. Nevertheless, assumption 1 can be ensured by design in settings where treatment timing can be explicitly randomized, such as our application in section IV. Moreover, it is frequently the case that the justification given for the validity of the parallel-trends assumption also justifies assumption 1. For example, Fadlon and Nielsen (2021, 12–13) write that the plausibility of the parallel-trends assumption in their context “relies on the notion that . . . the particular year at which the event occurs may be as good as random”; see, for example, Parker et al. (2013), Deshpande and Li (2019), and Nekoei and Seim (2023) for similar justifications.

It is also worth emphasizing that in nonexperimental contexts, the random-timing assumption may be more plausible if one restricts attention to units who are eventually treated. For example, Deshpande and Li (2019, 223) write that “some factors consistently predict the likelihood of a closing [i.e., the treatment] . . . However, no observable characteristic consistently predicts the timing of a closing conditional on closing . . . These results suggest that the timing of closings is effectively random even if the closings themselves are not.” Although in principle one can use DiD methods to exploit variation only among eventually treated units, units who are never treated are often included in DiD analyses to increase precision. In settings where the eventually treated units are more similar to each other than to the never-treated units, it therefore may be preferable to impose assumption 1 and use our efficient estimator than to use a DiD estimator that relies on parallel trends among never-treated units to increase efficiency. We also note that assumption 1 has testable implications, as we discuss in section II.H below, so researchers considering using our methodology in nonexperimental contexts can partially test the validity of assumption 1.

Analogously, Imbens (2004, 8) argues that while mean independence is technically weaker than full independence, arguments for the former often also justify the latter.

For example, the main specification in Bailey and Goodman-Bacon (2015) includes never-treated units, although the appendix shows results for an alternative specification that includes only eventually treated units, with substantially larger standard errors (contrast figs. 5 and E1).
Finally, we note that the validity of the parallel-trends assumption will typically be sensitive to functional form if treatment timing is not random (Roth and Sant’Anna 2023b). Empirical researchers should therefore be explicit about the justification for identification. If the parallel-trends assumption is justified on the basis of quasi-random treatment timing, then the methods developed in this paper can be used to obtain more precise estimates. On the other hand, if random treatment timing is not plausible, then methods that rely only on a parallel-trends assumption will be more appropriate. In this case, however, the researcher should provide a justification for why they expect parallel trends to hold specifically for the choice of functional form used in the analysis.

In addition to random treatment timing, we also assume that the treatment has no causal impact on the outcome in periods before it is implemented. This assumption is plausible in many contexts but may be violated if individuals learn of treatment status beforehand and adjust their behavior in anticipation (Abbring and van den Berg 2003; Lechner 2010; Malani and Reif 2015).7

**Assumption 2 (No anticipation).** For all \( i, Y_a(g) = Y_a(g') \) for all \( g, g' > t \).

Note that this assumption does not restrict the possible dynamic effects of treatment—that is, we allow for \( Y_a(g) \neq Y_a(g') \) whenever \( t \geq \min(g, g') \), so that treatment effects can arbitrarily depend on calendar time and the time that has elapsed since treatment. Rather, we require only that, say, a unit’s outcome in period 1 does not depend on whether it was ultimately treated in period 2 or period 3.8

**Example 1 (Special case: two periods).** Consider the special case of our model in which there are two periods \( (T = 2) \) and units are either treated in period 2 or never treated \( (G = \{2, \infty\}) \). Under random treatment timing and no anticipation, this special case is isomorphic to a cross-sectional experiment where the outcome \( Y_i = Y_a \) is the second-period outcome, the binary treatment \( D_i = 1[G_i = 2] \) is whether a unit is treated in period 2, and the covariate \( X_i = Y_{i1} = Y_{i\infty} \) is the pretreatment outcome (which by the no-anticipation assumption does not depend on treatment status). Covariate adjustment in cross-sectional randomized experiments has been studied by Freedman (2008a, 2008b), Lin (2013), and Li and Ding (2017), and our results will nest many of the existing results in the literature as a special case. The two-period special case also

---

7 If anticipatory behavior is possible only within \( m \) periods of treatment (e.g., because treatment is announced \( m \) periods in advance), the initial treatment can be redefined as \( G_i = m \).

8 Under the no-anticipation assumption, \( Y_i(g) \) can be interpreted as the outcome in period \( t \) from having been treated for \( \max(0, t - g) \) periods. We thank a referee for noting this interpretation.
allows us to study the canonical DiD estimator while avoiding complications discussed in the recent literature related to extending this estimator to the staggered case. We therefore come back to this example throughout the paper to provide intuition and connect our results to the previous literature.

Notation.—All expectations $\mathbb{E}[\cdot]$ and probability statements $\mathbb{P}(\cdot)$ are taken over the distribution of $G$ conditional on the potential outcomes and the number of units treated at each period, $(N_g)_{g \in G}$, although we suppress this conditioning for ease of notation. For a nonstochastic attribute $W_i$ (e.g., a function of the potential outcomes), we denote by $\mathbb{E}_f[W_i] = N^{-1} \sum_i W_i$ and $\text{Var}_f[W_i] = (N - 1)^{-1} \sum_i (W_i - \mathbb{E}_f[W_i])^2$ the finite-population expectation and variance, respectively, of $W_i$.

B. Target Parameters

In our staggered-treatment setting, the effect of being treated may depend on both the calendar time ($t$) and the time at which one was first treated ($g$). We therefore consider a large class of target parameters that allow researchers to highlight various dimensions of heterogeneous treatment effects across both calendar time and time since treatment.

Following Athey and Imbens (2022), we define $\tau_{t,gg'} = Y_i(g) - Y_i(g')$ to be the causal effect of switching the treatment date from $g'$ to $g$ on unit $i$’s outcome in period $t$. We define $\tau_{t,gg'} = N^{-1} \sum_t \tau_{t,gg'}$ to be the ATE of switching treatment from $g'$ to $g$ on outcomes at period $t$. We consider scalar estimands of the form

$$\theta = \sum_{t,gg'} a_{t,gg'} \tau_{t,gg'},$$

that is, weighted sums of the ATEs of switching from treatment $g'$ to $g$, with $a_{t,gg'} \in \mathbb{R}$ being arbitrary weights. Researchers will often be interested in weighted averages of the $\tau_{t,gg'}$, in which case the $a_{t,gg'}$ will sum to 1, although our results allow for arbitrary $a_{t,gg'}$. The results extend easily to vector-valued $\theta$’s where each component is of the form in the previous display; we focus on the scalar case for ease of notation. The no-anticipation assumption (assumption 2) implies that $\tau_{t,gg'} = 0$ if $t < \min(g, g')$, and so without loss of generality we make the normalization that $a_{t,gg'} = 0$ if $t < \min(g, g')$.

Example 1 (continued). In our simple two-period example, a natural target parameter is the ATE in period 2. This corresponds with setting $\theta = \tau_{2,2} = N^{-1} \sum_i Y_i(2) - Y_i(\infty)$.

9 This allows the possibility that, for instance, $\theta$ represents the difference between long-run and short-run effects, so that some of the $a_{t,gg'}$ are negative.
We now describe a variety of intuitive parameters that can be captured by this framework in the general staggered setting. Researchers are often interested in the effect of receiving treatment at a particular time relative to not receiving treatment at all. We define $\text{ATE}(t, g) := \tau_{t,g}$ to be the ATE on the outcome in period $t$ of being first treated at period $g$ relative to not being treated at all. The $\text{ATE}(t, g)$ is a close analogue to the cohort average treatment effects on the treated (ATTs) considered in Callaway and Sant’Anna (2021) and Sun and Abraham (2021). The main difference is that those papers do not assume random treatment timing, and thus they consider ATTs rather than ATEs.

In some cases, the $\text{ATE}(t, g)$ will be directly of interest and can be estimated in our framework. When the dimension of $t$ and $g$ is large, however, it may be desirable to aggregate the $\text{ATE}(t, g)$ both for ease of interpretability and to increase precision. Our framework incorporates a variety of possible summary measures that aggregate the $\text{ATE}(t, g)$ across different cohorts and time periods. We briefly discuss a few possible aggregations that may be relevant in empirical work, mirroring proposals for aggregating the ATT($t, g$) in Callaway and Sant’Anna (2021).

When researchers are interested in how the treatment effect evolves with respect to the time elapsed since treatment started, they may want to consider “event-study” parameters that aggregate the ATEs at a given lag $l$ since treatment ($l = 0, 1, \ldots$),

$$\theta_{l}^{ES} = \frac{1}{\sum_{g \cdot g+l \leq T} \sum_{g \cdot g+l \leq T} N_{g} \cdot \text{ATE}(g + l, g).}$$

Note that the instantaneous parameter $\theta_{0}^{ES}$ is analogous to the estimand considered in de Chaisemartin and D’Haultfoeuille (2020) in settings like ours where treatment is an absorbing state (although their framework also extends to the more general setting where treatment turns on and off).

In other situations, it may be of interest to understand how the treatment effect differs over calendar time (e.g., during a boom or bust economy) or by the time that treatment began. In such cases, the summary parameters

$$\theta_{t} = \frac{1}{\sum_{g \cdot g \leq t} \sum_{g \cdot g \leq t} N_{g} \cdot \text{ATE}(t, g) \quad \text{and}}$$

$$\theta_{g} = \frac{1}{T - g + 1} \sum_{t \cdot t \geq g} \text{ATE}(t, g),$$

which respectively aggregate the ATEs for a particular calendar time or treatment adoption cohort, may be relevant.

Finally, researchers may be interested in a single summary parameter for the effect of a treatment. In this case, it may be instructive to consider
a simple average of the \(\text{ATE}(t, g)\) (weighted by cohort size),

\[
\theta_{\text{simple}} = \frac{1}{\sum_i \sum_{g: g \leq t} N_g \sum_{g: g \leq t} N_g \text{ATE}(t, g)},
\]

or to consider a weighted average of the time or cohort effects,

\[
\theta_{\text{calendar}} = \frac{1}{T} \sum_i \theta_i, \quad \text{or} \quad \theta_{\text{cohort}} = \frac{1}{\sum_{g: g \leq \infty} N_g \sum_{g: g \leq \infty} N_g \theta_g}.
\]

Since the most appropriate parameter will depend on context, we consider a broad framework that allows for efficient estimation of all of these (and other) parameters.\(^{10}\)

C. Class of Estimators Considered

We now introduce the class of estimators we consider. Intuitively, these estimators start with a sample analogue to the target parameter and linearly adjust for differences in outcomes for units treated at different times in periods before either was treated.

Let \(\bar{Y}_g = N_g^{-1} \sum D_{tg} Y_t\) be the sample mean of the outcome for treatment group \(g\) in period \(t\), and let \(\hat{\tau}_{t,gg'} = \bar{Y}_g - \bar{Y}_{g'}\) be the sample analogue of \(\tau_{t,gg'}\). We define

\[
\hat{\theta}_0 = \sum_{t,gg'} a_{t,gg'} \hat{\tau}_{t,gg'},
\]

which replaces the population means in the definition of \(\theta\) with their sample analogues.

We consider estimators of the form

\[
\hat{\theta}_g = \hat{\theta}_0 - \hat{X}' \beta,
\]

where, intuitively, \(\hat{X}\) is a vector of differences in means that are guaranteed to be mean zero under the assumptions of random treatment timing and no anticipation. Formally, we consider \(M\)-dimensional vectors \(\hat{X}\) where each element of \(\hat{X}\) takes the form

\[
\hat{X}_j = \sum_{(t,gg') : g' > t} b_{t,gg'}^j \hat{\tau}_{t,gg'},
\]

where the \(b_{t,gg'}^j \in \mathbb{R}\) are arbitrary weights. There are many possible choices for the vector \(\hat{X}\) that satisfy these assumptions. For example, \(\hat{X}\) could be a vector where each component equals \(\hat{\tau}_{t,gg'}\) for a different combination of \((t, g, g')\) with \(t < g, g'\). Alternatively, \(\hat{X}\) could be a scalar that takes a weighted

\(^{10}\) We note that if \(\infty \not\in \mathcal{G}\), then \(\text{ATE}(t, g)\) is identified only for \(t < \max \mathcal{G}\). In this case, all of the sums above should be taken only over the \((t, g)\) pairs for which \(\text{ATE}(t, g)\) is identified.
Thus, for is the canonical DiD estimator, where \( o \) the simple DiM in period 2 and adjusts by \( o \) Sant Yi experiments considered in Lin (2013) and Li and Ding (2017), treating \( o \) to the set of linear covariate-adjusted estimators for cross-sectional mators. In this special case, the set of estimators of the form \( o \) \( Y_{12} \) as a fixed covariate.11

Example 1 (Continued). In our running two-period example, \( X = \hat{t}_{1,2e} \) corresponds with the pretreatment difference in sample means be- tween the units first treated at period 2 and the never-treated units. Thus, \( \hat{\theta}_1 = \hat{t}_{2,2e} - \hat{t}_{1,2e} = (\hat{Y}_{22} - \hat{Y}_{2e}) - (\hat{Y}_{12} - \hat{Y}_{1e}) \)

is the canonical DiD estimator, where \( \hat{Y}_g \) represents the sample mean of \( Y_{1e} \) for units with \( G = g \). Likewise, \( \hat{\theta}_0 \) is the simple difference in means (DiM) in period 2 (\( \hat{Y}_{22} - \hat{Y}_{2e} \)). More generally, the estimator \( \hat{\theta}^g \) takes the simple DiM in period 2 and adjusts by \( \beta \) times the DiM in period 1. Thus, for \( \beta \in (0, 1) \), \( \hat{\theta}^g \) is a weighted average of the DiM and DiD estimators. In this special case, the set of estimators of the form \( \hat{\theta}^g \) is equivalent to the set of linear covariate-adjusted estimators for cross-sectional experiments considered in Lin (2013) and Li and Ding (2017), treating \( Y_{1t} \) as a fixed covariate.11

Example 2 (Callaway and Sant’Anna 2021). For settings where there is a never-treated group (\( \infty \in \mathcal{G} \)), Callaway and Sant’Anna (2021) consider the estimator

\[ \hat{\tau}_{\infty}^{CS} = \hat{\tau}_{1,e} - \hat{\tau}_{\infty}^{g,1,e}, \]

that is, a DiD that compares outcomes between periods \( t \) and \( g - 1 \) for the cohort first treated in period \( g \) relative to the never-treated cohort. Observe that \( \hat{\tau}_{\infty}^{CS} \) can be viewed as an estimator of \( \text{ATE}(t, g) \) of the form given in equation (3), with \( X = \hat{t}_{g-1,e} \) and \( \beta = 1 \). Likewise, Callaway and Sant’Anna (2021) consider an estimator that aggregates the \( \hat{\tau}_{\infty}^{CS} \), say \( \hat{\tau}_{\omega}^{CS} = \sum_{i \in \mathcal{G}} \omega_{i,g} \hat{\tau}_{i,g,e} \), which can be viewed as an estimator of the parameter \( \hat{\theta}_\omega = \sum_{i \in \mathcal{G}} \omega_{i,g} \text{ATE}(t, g) \) of the form of equation (3) with \( X = \sum_{i \in \mathcal{G}} \omega_{i,g} \hat{t}_{g-1,g,e} \) and

11 Lin (2013) and Li and Ding (2017) consider estimators of the form \( \tau(\beta_0, \beta_1) = (Y_i - \beta_1(X_i - 1)) - (Y_i - \beta_0(X_i - 1)) \), where \( Y_i \) is the sample mean of the outcome \( Y_{1e} \) for units with treatment \( D = i, X_i \) is defined analogously, and \( \bar{X} \) is the unconditional mean of \( X \). Setting \( Y_1 = Y_{11}, X = Y_{11}, \) and \( D = 1[G = 2] \), it is straightforward to show that the estimator \( \tau(\beta_0, \beta_1) \) is equivalent to \( \hat{\theta}_\beta \) for \( \beta = (\bar{Y}_{1e}/N)\beta_0 + (\bar{Y}_{e}/N)\beta_1 \).
\[ \hat{\tau}_{CGS}^{CS2} = \frac{1}{\sum_{g' : g' > t} N_{g'}} \sum_{g' : g' > t} N_{g'} \hat{\tau}_{g'g} - \frac{1}{\sum_{g' : g' > t} N_{g'}} \sum_{g' : g' > t} N_{g'} \hat{\tau}_{g'-1,g}, \text{ for } t \geq g. \]

It is again apparent that this estimator can be written as an estimator of ATE\((t, g)\) of the form in equation (3), with \(X\) now corresponding with a weighted average of \(\hat{\tau}_{g'-1,g}\) and \(\beta\) again equal to 1.

**Example 3** (Sun and Abraham 2021). Sun and Abraham (2021) consider an estimator that is equivalent to that in Callaway and Sant’Anna (2021) in the case where there is a never-treated cohort. When there is no never-treated group, Sun and Abraham (2021) propose using the last cohort to be treated as the comparison. Formally, they consider the estimator of ATE\((t, g)\) of the form

\[ \hat{\tau}_{SA}^{CS2} = \hat{\tau}_{t,g_{\text{max}}} - \hat{\tau}_{g-1,g_{\text{max}}}, \]

where \(g_{\text{max}} = \max G\) is the last period in which units receive treatment. It is clear that \(\hat{\tau}_{SA}^{CS2}\) takes the form of equation (3), with \(X = \hat{\tau}_{g-1,g_{\text{max}}} \) and \(\beta = 1\). Weighted averages of the \(\hat{\tau}_{SA}^{CS2}\) can likewise be expressed in the form of equation (3), as with the Callaway and Sant’Anna (2021) estimators.

**Example 4** (de Chaisemartin and D’Haultfoeuille 2020). De Chaisemartin and D’Haultfoeuille (2020) propose an estimator of the instantaneous effect of a treatment. Although their estimator extends to settings where treatment turns on and off, in a setting like ours where treatment is an absorbing state, their estimator can be written as a linear combination of the \(\hat{\tau}_{CGS}^{CS2}\). In particular, their estimator is a weighted average of the Callaway and Sant’Anna (2021) estimators for the first period in which a unit was treated,

\[ \hat{\tau}_{CDH}^{CS2} = \frac{1}{\sum_{g : g \leq T} N_{g}} \sum_{g : g \leq T} N_{g} \hat{\tau}_{g_{\text{max}}}^{CS2}. \]

It is thus immediate from the previous examples that their estimator can also be written in the form of equation (3).

**Example 5** (TWFE models). Athey and Imbens (2022) consider the setting with \(G = \{1, \ldots, T, \infty\}\). Let \(A_{i,t} = 1[G_{i,t} \leq t]\) be an indicator for whether unit \(i\) is already treated by period \(t\). Athey and Imbens (2022, lemma 5) show that the coefficient on \(A_{i,t}\) from the TWFE specification

\[ \text{This could also be viewed as an estimator of the form of eq. (3) if } \hat{X} \text{ were a vector with each element corresponding with } \hat{\tau}_{g_{\text{max}}}, \text{ and the vector } \beta \text{ was a vector with elements corresponding with } a_{g_{\text{max}}}. \]
\[ Y_{it} = \alpha_i + \lambda_t + A_{it} \theta_{\text{TWFE}} + \epsilon_{it} \]  

(4)
can be decomposed as

\[ \hat{\theta}_{\text{TWFE}} = \sum_{t} \sum_{(g', g): \min(g, g') < t} \gamma_{t, gg'} \tilde{t}_{t, gg'} + \sum_{t} \sum_{(g', g): \min(g, g') > t} \tilde{\gamma}_{t, gg'} \tilde{t}_{t, gg'} \]  

(5)
for weights \( \gamma_{t, gg'} \), \( \tilde{\gamma}_{t, gg'} \) that depend only on the \( N_g \) and thus are nonstochastic in our framework. Thus, \( \hat{\theta}_{\text{TWFE}} \) can be viewed as an estimator of the form of equation (3) for the parameter \( \theta_{\text{TWFE}} = \sum_t \sum_{(g', g): \min(g, g') < t} \gamma_{t, gg'} \tau_{t, gg'} , \) with \( \tilde{\tau} = -\sum_t \sum_{(g', g): \min(g, g') > t} \tilde{\gamma}_{t, gg'} \tilde{t}_{t, gg'} \) and \( \beta = 1 \). As noted in Athey and Imbens (2022) and other papers, however, the parameter \( \theta_{\text{TWFE}} \) may be difficult to interpret under treatment effect heterogeneity, since the weights \( \gamma_{t, gg'} \) are not guided by economic reasoning, and moreover, some of the \( \gamma_{t, gg'} \) may be negative, so that \( \theta_{\text{TWFE}} \) is not a convex-weighted average of causal effects.

We note that, in principle, one can also use a vector-valued \( \tilde{\tau} \) that stacks the \( \tilde{\tau} \) values used by multiple estimators. For example, one could set

\( \tilde{\tau} = (\tilde{\tau}^{CS}, \tilde{\tau}^{CS2})' \),

which combines the scalar values of \( \tilde{\tau} \) used for the two variants of the Callaway and Sant’Anna (2021) estimator. Then, \( \hat{\theta}_{(0,1)} \) would correspond to \( \tilde{\tau}^{CS} \), while \( \hat{\theta}_{(0,1)} \) would correspond to \( \tilde{\tau}^{CS2} \), thus nesting both estimators in the class of estimators of the form \( \hat{\theta}_{\beta} \). One could likewise stack the \( \tilde{\tau} \)'s associated with other DiD-related estimators. We stress, though, that the notion of efficiency that we derive below will be for the class of estimators using a specific vector \( \tilde{\tau} \); see remarks 4 and 5 below for additional discussion.

D. Efficient “Oracle” Estimation

We now consider the problem of finding the best estimator \( \hat{\theta}_{\beta} \) of the form introduced in equation (3). We first show that \( \hat{\theta}_{\beta} \) is unbiased for all \( \beta \) and then solve for the \( \beta^{*} \) that minimizes the variance.

Notation.—We begin by introducing some notation that will be useful for presenting our results. Recall that the sample treatment effect estimates \( \tilde{t}_{t, gg'} \) are themselves differences in sample means, \( \tilde{t}_{t, gg'} = \tilde{Y}_{tg} - \tilde{Y}_{tg'} \). It follows that we can write

\[ \hat{\theta}_{\beta} = \sum_{g} A_{\beta g} \tilde{Y}_{g} \quad \text{and} \quad \tilde{\tau} = \sum_{g} A_{\beta g} \tilde{Y}_{g} \]

for appropriately defined matrices \( A_{\beta g} \) and \( A_{\beta g} \) of dimensions \( 1 \times T \) and \( M \times T \), respectively, where \( \tilde{Y}_{g} = (\tilde{Y}_{1g}, ..., \tilde{Y}_{Tg})' \). Additionally, let \( S_g = \text{Var} [Y_i(g)] \) be the finite-population variance of \( Y_i(g) \), and let

\[ S_{gg'} = (N - 1)^{-1} \sum_i (Y_i(g) - \bar{Y}_g)(Y_i(g') - \bar{Y}_g') \]

be the finite-population covariance between \( Y_i(g) \) and \( Y_i(g') \).
Our first result is that all estimators of the form $\hat{\theta}_\beta$ are unbiased, regardless of $\beta$.

**Lemma 2.1 ($\hat{\theta}_\beta$ unbiased).** Under assumptions 1 and 2, $E[\hat{\theta}_\beta] = \theta$ for any $\beta \in \mathbb{R}^M$.

See remark 6 below (sec. II.E.2) for a discussion of the bias that arises when assumption 1 fails.

We next turn our attention to finding the value $\beta^*$ that minimizes the variance.

**Proposition 2.1.** Under assumptions 1 and 2, the variance of $\hat{\theta}_\beta$ is uniquely minimized at $\beta^* = Var[\hat{\beta}]^{-1} Cov[\hat{\beta}, \hat{\theta}_0]$, provided that $Var[\hat{\beta}]$ is positive definite. Further, the variances and covariances in the expression for $\beta^*$ are given by

$$Var\left(\begin{pmatrix} \hat{\theta}_0 \\ \hat{\beta} \end{pmatrix}\right) = \left(\sum_x N_x^{-1} A_{0x} S_{xx} A_{0x}^\prime - N^{-1} S_x, \sum_x N_x^{-1} A_{0x} S_{x0} A_{0x}^\prime \right) = \left(\begin{pmatrix} V_{\theta_0} & V_{\theta_0, \hat{\beta}} \\ V_{\theta_0, \hat{\beta}} & V_{\hat{\beta}} \end{pmatrix}\right),$$

where $S_x = Var[\sum_x A_{0x} Y_i(g)]$. The efficient estimator has variance given by $Var[\hat{\theta}_{\beta^*}] = V_{\theta_0} - (\beta^*)' V_{\hat{\beta}}^{-1} (\beta^*)$.

Equation (6) shows that the variance-minimizing $\beta^*$ is the best linear predictor of $\hat{\theta}_0$, given $\hat{X}$. This formalizes the intuition that it is efficient to place more weight on pretreatment differences in outcomes the more strongly they correlate with the posttreatment differences in outcomes.

**Example 1 (continued).** In our ongoing two-period example, the efficient estimator $\hat{\theta}_{\beta^*}$ derived in proposition 2.1 is equivalent to the efficient estimator for cross-sectional randomized experiments in Lin (2013) and Li and Ding (2017). The optimal coefficient $\beta^*$ is equal to $(N_1/N)\beta_2 + (N_2/N)\beta_3$, where $\beta_2$ is the coefficient on $Y_{1i}$ from a regression of $Y_{1i}(g)$ on $Y_{1i}$ and a constant. Intuitively, this estimator puts more weight on the pretreatment outcomes (i.e., $\beta^*$ is larger) the more predictive is the first-period outcome $Y_{1i}$ of the second-period potential outcomes. In the special case where the coefficients on lagged outcomes are equal to 1, the canonical DiD estimator is optimal, whereas the simple DiM estimator is optimal when the coefficients on lagged outcome are zero. For values of $\beta^* \in (0, 1)$, the efficient estimator can be viewed as a weighted average of the DiD and DiM estimators.

**E. Properties of the Plug-In Estimator**

Proposition 2.1 solves for the $\beta^*$ that minimizes the variance of $\hat{\theta}_\beta$. However, the efficient estimator $\hat{\theta}_{\beta^*}$ is not of practical use, since the “oracle” coefficient $\beta^*$ depends on the covariances of the potential outcomes, $S_x$. 
which are typically not known in practice. Mirroring Lin (2013) for cross-sectional randomized experiments, we now show that $\beta^*$ can be approximated by a plug-in estimate $\hat{\beta}^*$ and that the resulting estimator $\hat{v}^*$ has properties similar to those of the “oracle” estimator $\hat{\theta}_{\beta^*}$ when $N$ is large.

1. Definition of the Plug-In Estimator

To formally define the plug-in estimator, let

$$\hat{S}_g = \frac{1}{N_g - 1} \sum_i D_g(y_i(g) - \hat{Y}_g)(Y_i(g) - \hat{Y}_g)'$$

be the sample analogue to $S_g$, and let $\hat{V}_X$, $\hat{V}_{X,k}$, $\hat{V}_{X,h}$ be the respective analogues to $V_X$, $V_{X,k}$, and $V_{X,h}$ that replace $S_g$ with $\hat{S}_g$ in the definitions. We then define the plug-in coefficient

$$\hat{\beta}^* = \hat{V}_X^{-1} \hat{V}_{X,h},$$

and consider the properties of the plug-in efficient estimator $\hat{v}^*$.

**Example 1** (continued). In our ongoing two-period example, which we have shown is analogous to a cross-sectional randomized experiment, the plug-in estimator $\hat{\theta}_{\beta^*}$ is equivalent to the efficient plug-in estimator for cross-sectional experiments considered in Lin (2013). As in Lin (2013), $\hat{\theta}_{\beta^*}$ can be represented as the coefficient on $D_i$ in the interacted ordinary least squares (OLS) regression,

$$Y_{\omega} = \beta_0 + \beta_1 D_i + \beta_2 \hat{Y}_{i1} + \beta_3 D_i \times \hat{Y}_{i1} + \epsilon_i,$$

where $\hat{Y}_{i1}$ is the de-meaned value of $Y_{i1}$. Intuitively, this fully interacted specification fits one linear model to estimate the mean of $Y_{\omega}(2)$ and another to estimate $Y_{\omega}(\infty)$ and then computes the difference and thus is an augmented inverse-propensity-weighted (AIPW) estimator with a linear model for the conditional expectation functions and a constant propensity score (Glynn and Quinn 2010).

**Remark 3** (Connection to McKenzie 2012). McKenzie (2012) proposes using an estimator similar to the plug-in efficient estimator in the two-period setting considered in our ongoing example. Building on results in Frison and Pocock (1992), he proposes using the coefficient $\gamma_1$ from the OLS regression

$$Y_{\omega} = \gamma_0 + \gamma_1 D_i + \gamma_2 \hat{Y}_{i1} + \epsilon_i,$$

which is sometimes referred to as the analysis of covariance (ANCOVA I).

We are not aware of a representation of the plug-in efficient estimator as the coefficient from an OLS regression in the more general, staggered case.
estimator in equation (7), sometimes referred to as ANCOVA II, in that it omits the interaction term $D_i \hat{Y}_i$. Treating $\hat{Y}_i$ as a fixed pretreatment covariate, the coefficient $\hat{g}_1$ from equation (8) is equivalent to the estimator studied in Freedman (2008a, 2008b). The results in Lin (2013) therefore imply that McKenzie’s (2012) estimator will have the same asymptotic efficiency as $\hat{\theta}_v$ under constant treatment effects. Intuitively, this is because the coefficient on the interaction term in equation (7) converges in probability to 0. However, the results in Freedman (2008a, 2008b) imply that under heterogeneous treatment effects, McKenzie’s (2012) estimator may even be less efficient than the simple DiM $\hat{\theta}_v$, which in turn is (weakly) less efficient than $\hat{\theta}_v^*$.14

2. Asymptotic Properties of the Plug-In Estimator

We now show that in large populations, the plug-in efficient estimator $\hat{\theta}_v^*$ is asymptotically unbiased for $\theta$ and has the same asymptotic variance as the oracle estimator $\hat{\theta}_v^*$. To derive the properties of the plug-in efficient estimator in large finite populations, we consider a sequence of finite populations of increasing sizes, as in Lin (2013) and Li and Ding (2017), among other papers. More formally, we consider sequences of populations indexed by $m$ where the number of observations first treated at period $g$, $N_{g,m}$, diverges for all $g \in \mathcal{G}$. For ease of notation, as in the aforementioned papers we leave the index $m$ implicit in our notation for the remainder of the paper. We assume that the sequence of populations satisfies the following regularity conditions:

Assumption 3.

i) For all $g \in \mathcal{G}$, $N_g/N \to p_g \in (0, 1)$.

ii) For all $g$ and $g'$, $S_g$ and $S_{gg'}$ have limiting values denoted $S_g^*$ and $S_{gg'}^*$, respectively, with $S_g^*$ positive definite.

iii) $\max_{i,g} \|Y_i(g) - \mathbb{E}_f[Y_i(g)]\|^2/N \to 0$.

Part i imposes that the fraction of units first treated at period $g \in \mathcal{G}$ converges to a constant bounded between 0 and 1. Part ii requires that the variances and covariances of the potential outcomes converge to a constant. Part iii requires that no single observation dominates the finite-population variance of the potential outcomes and is thus analogous to the familiar Lindeberg condition in sampling contexts.

With these assumptions in hand, we are able to formally characterize the asymptotic distribution of the plug-in efficient estimator. The following

14 Relatedly, Yang and Tsiatis (2001), Funatogawa, Funatogawa, and Shyr (2011), Wan (2020), and Negi and Wooldridge (2021) show that $\hat{\theta}_v$ from eq. (7) is asymptotically at least as efficient as $\hat{g}_1$ from eq. (8) in sampling-based models similar to our ongoing example.
result shows that $\hat{\theta}_\beta^*$ is asymptotically unbiased and normally distributed, with the same asymptotic variance as the “oracle” efficient estimator $\hat{\theta}_\beta$. The proof exploits the general finite-population central limit theorem in Li and Ding (2017).

**Proposition 2.2.** Under assumptions 1, 2, and 3,

$$\sqrt{N}(\hat{\theta}_\beta^* - \theta) \to_d N(0, \sigma_n^2),$$

where $\sigma_n^2 = \lim_{N \to \infty} N \text{Var}[\hat{\theta}_\beta^*]$.

**Remark 4 (Connection to semiparametric efficiency).** Proposition 2.2 shows that the plug-in estimator $\hat{\theta}_\beta^*$ achieves the same asymptotic variance as $\hat{\theta}_\beta$, the most efficient estimator in the class $\hat{\theta}_\beta$. We note that the asymptotic variance of the best estimator in this class is distinct from the semiparametric efficiency bound in superpopulation frameworks (e.g., Hahn 1998; Sant’Anna and Zhao 2020). We are not aware of any results on semiparametric efficiency in design-based frameworks such as ours, nor are we aware of any results on the semiparametric efficiency bound for panel data settings with staggered treatment timing—although both of these strike us as interesting directions for future research. Existing results do suggest a connection between our notion of efficiency and semiparametric efficiency in our ongoing two-period example, however. Negi and Wooldridge (2021) study covariate adjustment in cross-sectional randomized experiments from a superpopulation perspective and show that Lin’s (2013) estimator (which they refer to as full regression adjustment [FRA]) achieves the semiparametric efficiency bound when the conditional expectation of the potential outcomes is linear in the observed covariates. Since our estimator is equal to FRA in our running two-period example (viewing $Y_1$ as the pretreatment covariate), this implies that $\hat{\theta}_\beta^*$ is semiparametric efficient (from the superpopulation perspective) when the conditional expectations of the second-period potential outcomes are linear in the pretreatment outcome.

**Remark 5 (On the choice of $\hat{X}$).** We note that $\text{Var}[\hat{\theta}_\beta^*] = V_{\theta_{\beta^*}} - (\beta^*)' V_X^{-1} (\beta^*)$ can be viewed as the variance of the residual after linearly projecting $\theta_{\beta}$ onto $\hat{X}$. Thus, the asymptotic variance of the plug-in efficient estimator will be smaller if $\hat{X}$ is more predictive of the estimation error in the simple DiM estimator $\hat{\theta}_0$. It thus may be tempting to set $\hat{X}$ to be a vector including all possible comparisons of cohorts in periods before they were treated in order to minimize the asymptotic variance. This may not improve finite-sample performance, however, since the asymptotics considered in proposition 2.2 assume that the number of observations $N$ is substantially larger than the dimension of $\hat{X}$ and thus may not approximate the finite-sample performance when $\text{dim}(\hat{X})$ is large. In particular, using a too-high-dimensional $\hat{X}$ may lead to an “overfitting” problem analogous to controlling for too many pretreatment variables in a cross-sectional experiment (see sec. III for an example of this phenomenon). Lei and Ding
(2021) study covariate adjustment with a diverging number of covariates in cross-sectional randomized experiments, and find that (under certain regularity conditions), linear covariate adjustment works well when the dimension of the covariates is small relative to $N^{-1/2}$. We suspect that a similar heuristic applies to the choice of the dimension of $\hat{X}$, although we leave a formal analysis under diverging covariates to future work. In our Monte Carlo simulations below, we find good performance for the scalar $\hat{X}$ such that $\beta = 1$ corresponds to the Callaway and Sant’Anna (2021) estimator, and thus consider this a reasonable default for practitioners implementing our method.

Remark 6 (Bias under nonrandom timing). Lemma 2.1 shows that the oracle efficient estimator $\hat{\theta}_{\beta^*}$ is unbiased under random treatment timing and no anticipation. If, however, the random-treatment-timing assumption is violated, then the efficient estimator may be biased, whereas the DiD estimator ($\beta = 1$) may still be unbiased under a parallel-trends assumption. We note that $\hat{\theta}_{\beta^*} - \hat{\theta}_1 = (\beta^* - 1)\hat{X}$, and thus $\mathbb{E}[\hat{\theta}_{\beta^*} - \hat{\theta}_1] = (\beta^* - 1)\mathbb{E}[\hat{X}]$ (assuming that $\hat{X}$ is scalar, for simplicity). Hence, when DiD is unbiased but random treatment timing is violated, the bias of the oracle efficient estimator will be larger (i) the farther is $\beta^*$ from 1 and (ii) the larger is $\mathbb{E}[\hat{X}]$, that is, the more imbalance there is in the pretreatment outcome. We note, however, that when treatment timing is nonrandom, the DiD estimator will often be biased as well, for example, when treatment is randomly assigned conditional on lagged outcomes (Angrist and Pischke 2009; Ding and Li 2019).

F. Inference

We now introduce two methods for inference on $\theta$, the first using conventional $t$-based CIs and the second using FRTs.

1. $t$-Based CIs

To construct CIs using the asymptotic normal distribution derived in proposition 2.2, one requires an estimate of the variance $\sigma_{\theta^*}^2$. We first show that a simple Neyman-style variance estimator is conservative under treatment effect heterogeneity, as is common in finite-population

---

15 Future work might also consider an estimator that uses a high-dimensional $\hat{X}$ but considers some form of regularization on the coefficient $\hat{\beta}$.

16 As noted above, in the simple two-period example, the plug-in efficient estimator is equivalent to an AIPW estimator with a constant propensity score and a linear model for the conditional expectation function. Thus, from a superpopulation perspective, the plug-in efficient estimator would be consistent under the conditional unconfoundedness assumption, $1[G_i = 2] \perp Y_{it} | Y_{it}$, when the conditional expectation functions are linear (see, e.g., Hahn 1998). Formalizing this type of robustness in our design-based framework and extending it to settings with staggered treatment timing strikes us an interesting direction for future work.
settings. We then introduce a less conservative refinement to this estimator that adjusts for the part of the heterogeneity explained by $X$.

Recall that $\sigma_r^2 = \lim_{N \to \infty} N \text{Var}[\hat{\theta}_r^2]$. Examining the expression for $\text{Var}[\hat{\theta}_r^2]$ given in proposition 2.1, we see that all of the components of the variance can be replaced with sample analogues except for the $-S_g$ term. This term corresponds with the variance of treatment effects and is not consistently estimable, since it depends on covariances between potential outcomes under treatments $g$ and $g'$ that are never observed simultaneously. This motivates the use of the Neyman-style variance that ignores the $-S_g$ term and replaces the variances $S_g$ with their sample analogues $\hat{S}_g$,

$$\hat{\sigma}_r^2 = \left( \sum_{g} \frac{N}{N_g} A_{g,g} \hat{S}_g A'_{g,g} \right) \left( \sum_{g} \frac{N}{N_g} A_{0,g} \hat{S}_g A'_{0,g} \right) \left( \sum_{g} \frac{N}{N_g} A_{g,g} \hat{S}_g A'_{0,g} \right)' \left( \sum_{g} \frac{N}{N_g} A_{0,g} \hat{S}_g A'_{0,g} \right)^{-1}.$$

Since $\hat{S}_g \to_p S_g^*$ (see lemma A.2), it is immediate that the estimator $\hat{\sigma}_r^2$ converges to an upper bound on the asymptotic variance $\sigma_r^2$, although the upper bound is conservative if there are heterogeneous treatment effects such that $S_g^* = \lim_{N \to \infty} S_g > 0$.

**Lemma 2.2.** Under assumptions 1, 2, and 3, $\hat{\sigma}_r^2 \to_p \sigma_r^2 + S_g^* \geq \sigma_r^2$.

The estimator $\hat{\sigma}_r^2$ can be improved by using outcomes from earlier periods. The refined estimator intuitively lower-bounds the heterogeneity in treatment effects by the part of the heterogeneity that is explained by the outcomes in earlier periods. The construction of this refined estimator mirrors the refinements using fixed covariates in randomized experiments considered in Lin (2013) and Abadie et al. (2020), with lagged outcomes playing a role similar to that of the fixed covariates. To avoid technical clutter, we defer the construction of the refined variance estimator to appendix A.1 and merely state the sense in which the refined estimator improves upon the Neyman-style estimator introduced above.

**Lemma 2.3.** The refined estimator $\hat{\sigma}_{rs}^2$, defined in lemma A.4, satisfies $\hat{\sigma}_{rs}^2 \to_p \sigma_r^2 + S_g^*$, where $0 \leq S_g^* \leq \hat{S}_g^*$, so that $\hat{\sigma}_{rs}^2$ is asymptotically (weakly) less conservative than $\hat{\sigma}_r^2$.

It is then immediate that $\text{CI}_{rs} = \hat{\theta}_r^* \pm z_{1-\alpha/2} \cdot \hat{se}$ is a valid $(1 - \alpha)$-level CI for $\theta$, where $\hat{se} = \hat{\sigma}_{rs}/\sqrt{n}$ is the standard error and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the normal distribution.

2. **FRTs**

An alternative approach to inference uses FRTs, otherwise known as permutation tests. We will show that an FRT using a studentized version of
the efficient estimator has the dual advantages that it (i) has exact size under the sharp null of no treatment effects for all units and (ii) is asymptotically valid for the weak null that $\theta = 0$.

To derive the FRT, recall that the observed data are $(Y, G)$, where $Y$ collects all of the $Y_i$ and $G = (G_1, ..., G_N)'$. Let $T = T(Y, G)$ denote a statistic of the data, and let $T_\pi = T(Y, G_\pi)$ be the statistic using the transformed data in which $G$ is replaced with a permutation $G_\pi$. An FRT computes the $p$-value

$$p_{\text{FRT}} = P_{\pi \sim U(\Pi)}(T_\pi \geq T(Y, G)),$$

where the probability is taken over the uniform distribution on the set of permutations $\Pi$. Under the sharp null hypothesis that $Y_i(g) = Y_i(g')$ for all $i$, $g$, $g'$, the distribution of $T_\pi$ is the same as the distribution as $T(Y, G)$, and thus by standard arguments the FRT is exact in finite samples (see, e.g., Imbens and Rubin 2015).

The sharp null hypothesis of no treatment effect will often be too restrictive in practice, however, as we may be more interested in the hypothesis that the average effect is zero, that is, $H_0 : \theta = 0$. Unfortunately, in general FRTs may not have correct size for such weak null hypotheses, even asymptotically (Wu and Ding 2021).

We now show, however, that when the FRT is based on the studentized statistic $T(Y, G) = \hat{\theta}^{**}/\hat{se}$, it has asymptotically correct size under the weak null. In fact, we show that asymptotically the FRT is equivalent to testing that 0 falls within the $t$-based CI derived in the previous section. Thus, this FRT based on the studentized statistic is in some sense the “best of both worlds” of Fisherian and Neymanian inference, in that it has exact size under the sharp null hypothesis while having asymptotically correct size under the weak null.

The following regularity condition imposes that the means of the potential outcomes have limits and that their fourth moment is bounded.

Assumption 4. Suppose that for all $g$, $\lim_{N \to \infty} \mathbb{E} [Y_i(g)] = \mu_g < \infty$ and that there exists $L < \infty$ such that $N^{-1} \Sigma_i \|Y_i(g) - \mathbb{E}[Y_i(g)]\|^4 < L$ for all $N$.

With this assumption in hand, we can make precise the sense in which the FRT is asymptotically valid under the weak null.

Proposition 2.3. Suppose that assumptions 1–4 hold. Let $t_\pi = (\hat{\theta}^{**}/\hat{se})_\pi$ be the studentized $t$-statistic under permutation $\pi$. Then $t_\pi \to_d N(0, 1)$, $P_\infty$-almost surely. Hence, if $p_{\text{FRT}}$ is the $p$-value from the FRT associated with $|t_\pi|$, then under $H_0 : \theta = 0$,

$^{17}$ Formally, a permutation $\pi$ is a bijective map from $[1, ..., N]$ onto itself, and $G_\pi = (G_{\pi(i)}, ..., G_{\pi(N)})'$.

$^{18}$ It is often difficult to calculate the $p$-value over all permutations exactly, so the $p$-value is approximated via simulation. We use 500 simulation draws in our simulations and 5,000 draws in the empirical application.
\[
\lim_{N \to \infty} P(p_{\text{FRT}} \leq \alpha) \leq \alpha,
\]

\(P\) -almost surely, with equality if and only if \(S^*_\beta = 0\).

Proposition 2.3 implies that the FRT using the studentized version of the efficient estimator asymptotically controls size under the weak null of zero ATE. Indeed, the proposition implies that the FRT is asymptotically equivalent to the test that the \(t\) -based CI** includes 0. Proposition 2.3 extends the results in Wu and Ding (2021) and Zhao and Ding (2021), who consider permutation tests based on a studentized statistic in cross-sectional randomized experiments.\(^{19}\) Given the desirable properties of the FRT under both the sharp and weak null hypotheses, we recommend that researchers report \(p\) -values from the FRT alongside the usual \(t\) -based CIs.

**G. Implications for Existing Estimators**

We now discuss the implications of our results for estimators previously proposed in the literature. We have shown that in the simple two-period case considered in example 1, the canonical DiD corresponds with \(\hat{\theta}_1\). Likewise, in the staggered case, we showed in examples 2–4 that the estimators of Callaway and Sant’Anna (2021), Sun and Abraham (2021), and de Chaisemartin and D’Haultfoeuille (2020) correspond with the estimator \(\hat{\theta}_1\) for an appropriately defined estimand and \(\hat{X}\). Our results thus imply that, unless \(\beta^* = 1\), the estimator \(\hat{\theta}_{\beta^*}\) is unbiased for the same estimand and has strictly lower variance under (quasi-)random treatment timing. Since the optimal \(\beta^*\) depends on the potential outcomes, we do not generically expect \(\beta^* = 1\), and thus the previously proposed estimators will generically be dominated in terms of efficiency. Although the optimal \(\beta^*\) will typically not be known, our results imply that the plug-in estimator \(\hat{\theta}_{\beta^*}\) will have similar properties in large populations and thus will be more efficient than the previously proposed estimators in large populations under (quasi-)random treatment timing. We thus recommend the plug-in efficient estimator in settings where the parallel-trends assumption is justified with random treatment timing.

We note, however, that the estimators in the aforementioned papers are valid for the ATT in settings where only the parallel-trends assumption holds but there is not random treatment timing, whereas the validity of the efficient estimator depends on random treatment timing (see remark 2

\(^{19}\) Permutation tests based on a studentized statistic have been considered in other contexts as well, e.g., Janssen (1997); Chung and Romano (2013, 2016); DiCiccio and Romano (2017); Bugni, Canay, and Shaikh (2018); MacKinnon and Webb (2020); and Bai, Romano, and Shaikh (2022).
above). 20 Although in some settings the parallel-trends assumption is justified by arguing that treatment is (quasi-)randomly assigned, in some observational settings the researcher may be more comfortable imposing parallel trends than quasi-random treatment timing. We thus view the plug-in efficient estimator to be complementary to the estimators considered in previous work, since it is more efficient under stricter assumptions that will not hold in all cases of interest.

H. Extensions and Practical Considerations

We now discuss several extensions and practical considerations that may be useful for applying our methods.

Remark 7 (Testing the randomization assumption). It may often be desirable to test the assumption of (quasi-)random treatment timing, especially in nonexperimental settings where random timing cannot be ensured by design. We briefly describe three approaches. First, since consistency of the efficient estimator depends on the assumption that $E [\hat{X}] = 0$, a natural falsification test is to test whether $\hat{X}$ is significantly different from zero—that is, are there significant differences in pretreatment means between cohorts treated at different times? It is straightforward to conduct a test of the null that $E [\hat{X}] = 0$ using a one-sample $t$-test (with a sample analogue to the variance given in proposition 2.1) or using an FRT. Second, an intuitive approach that mirrors the common practice of testing for preexisting trends is to estimate an event study, treating the initial time of treatment as $G_i - k$ for some $k > 0$, and then test whether the dynamic effects corresponding with the leads $1, \ldots, k$ are different from zero. 21 Third, as is common in randomized controlled trials, researchers can test for covariate balance between units treated at different times. For example, Deshpande and Li (2019) show that observable characteristics do not predict the timing of social security office closings. We illustrate how these types of tests can be used in our application below. Such tests can be a useful test of the plausibility of the randomization assumption and can help to identify cases where it is clearly violated. We caution, however, that as with tests of preexisting trends (see Roth 2022), such falsification tests may have limited power to detect violations of the randomization assumption, and relying on them can introduce distortions from pretesting. Thus, it is best to additionally motivate the randomization assumption on the basis of context-specific knowledge.

20 The estimator of de Chaisemartin and D’Haultfoeuille (2020) can also be applied in settings where treatment turns on and off over time.

21 Given that the efficient estimator differs from the usual DiD estimator, note that this test differs from the common pretest for preexisting trends.
Remark 8 (Conditional random treatment timing). For simplicity, we have considered the case of unconditional random treatment timing. In some experiments, the treatment timing may be randomized among units with some shared observable characteristics (e.g., counties within a state). In this case, the methodology described above can be applied within each randomization stratum, and the stratum-level estimates can be pooled to form aggregate estimates for the population. Likewise, in quasi-experimental contexts, the assumption of quasi-random treatment timing may be more plausible among subgroups of the population (e.g., within units of the same gender and education status) or among groups of units that were treated at similar times (e.g., within a decade). The units can then be partitioned into strata based on discrete observable characteristics, and the analysis we describe can be conducted within each stratum. Extending our results to allow for randomization conditional on a continuous characteristic is an interesting topic for future work.

Remark 9 (Clustered treatment assignment). Likewise, in some settings there may be clustered assignment of treatment timing—for example, treatment is assigned to families $f$, and all units $i$ in family $f$ are first treated at the same time. This violates assumption 1, since not all vectors of treatment timing are equally likely. However, note that any average treatment contrast at the individual level, for example, $(1/N) \sum_i Y_i(g) - Y_i(g')$, can be written as an average contrast of a transformed family-level outcome, for example, $(1/F) \sum_f \tilde{Y}_f(g) - \tilde{Y}_f(g')$, where $\tilde{Y}_f(g) = (F/N) \sum_{i \in f} Y_i(g)$. Thus, clustered assignment can easily be handled in our framework by analyzing the transformed data at the cluster level.

Remark 10 (Fixed pretreatment covariates). In some settings, researchers may also have access to fixed pretreatment covariates $W_i$. Differences in the mean of $W_i$ between adoption cohorts can then be added to the vector $\hat{X}$ to further increase precision.

III. Monte Carlo Results

We present two sets of Monte Carlo results. In section III.A, we conduct simulations in a stylized two-period setting matching our ongoing example to illustrate how the plug-in efficient estimator compares to the classical DiD and simple DiM estimators. Section III.B presents a more realistic set of simulations with staggered treatment timing that is calibrated to our application, comparing the plug-efficient estimator to recent DiD-based estimators proposed for the staggered-treatment case.

22 The FRTs can likewise be modified to consider permutations that permute assignments only within randomization strata.
A. Two-Period Simulations

Specification.—We follow the model in example 1 in which there are two periods \((t = 1, 2)\) and units are treated in period 2 or never treated \((G = \{2, \infty\})\). We generate the potential outcomes as follows. For each unit \(i\) in the population, we draw the never-treated potential outcomes \(Y_i(\infty) = (Y_i(1), Y_i(2))\) from an \(N(0, \Sigma_r)\) distribution, where \(\Sigma_r\) has 1’s on the diagonal and \(r\) on the off diagonal. The parameter \(r\) is the correlation between the untreated potential outcomes in periods \(t = 1, 2\).

We then set \(Y_i(2) = Y_i(\infty) + \tau_i\), where \(\tau_i = \gamma(Y_i(\infty) - E[Y_i(\infty)])\). The parameter \(\gamma\) governs the degree of heterogeneity of treatment effects: if \(\gamma = 0\), then there is no treatment effect heterogeneity, whereas if \(\gamma\) is positive then individuals with larger untreated outcomes in period 2 have larger treatment effects. We center by \(E[Y_i(\infty)]\) so that the treatment effects are on average. We generate the potential outcomes once and treat the population as fixed throughout our simulations. Our simulation draws then differ on the basis of the draw of the treatment assignment vector. For simplicity, we set \(N_2 = N_\infty = N/2\), and in each simulation draw, we randomly select which units are treated in period 2 or not. We conduct 1,000 simulations for all combinations of \(N_2 \in \{25, 1000\}\), \(\rho \in \{0, 0.5, 0.99\}\), and \(\gamma \in \{0, 0.5\}\).

Results.—Table 1 shows the bias, standard deviation, and coverage of 95% CIs for the plug-in efficient estimator \(\hat{\theta}_{\beta^*}\), the DiD estimator \(\hat{\theta}_{DiD}\), and the simple DiM estimator \(\hat{\theta}_{DiM}\). It also shows the size (null rejection probability) of the FRT using a studentized statistic introduced in section II.F. The CIs are constructed as \(\hat{\theta}_{\beta^*} \pm 1.96\hat{\sigma}_{\beta^*}/\sqrt{N}\) for the plug-in efficient estimator and analogously for the other estimators.\(^{23}\) For all specifications and estimators, the estimated bias is small, and coverage is close to the nominal level. Table 2 facilitates comparison of the standard deviations of the different estimators by showing the ratio relative to the plug-in estimator. The standard deviation of the plug-in efficient estimator is weakly smaller than that of either DiD or DiM in nearly all cases and is never more than 2% larger than that of either DiD or DiM. The standard deviation of the plug-in efficient estimator is similar to that of DiD when autocorrelation of \(Y(\infty)\) is high \((\rho = 0.99)\) and there is no heterogeneity of treatment effects \((\gamma = 0)\), so that \(\beta^* \approx 1\) and thus DiD is (nearly) optimal in the class we consider. Likewise, it is similar to that of DiM when there is no autocorrelation \((\rho = 0)\) and there is no treatment effect heterogeneity \((\gamma = 0)\); thus, \(\beta^* \approx 0\) and so DiM is (nearly) optimal in the class we consider. The plug-in efficient estimator is substantially more precise than DiD and DiM in

\(^{23}\) For \(\hat{\theta}_{\beta^*}\), we use an analogue to \(\hat{\sigma}_{\beta^*}\), except that the unrefined estimate \(\hat{\sigma}_{\beta^*}\) is replaced with the sample analogue to the expression for \(\text{Var}[^{\hat{\theta}_{\beta^*}}]\) implied by proposition 2.1.
| Specification | Bias | SD | Coverage | FRT Size |
|---------------|------|----|----------|----------|
| N1 | N0 | \(\rho\) | \(\gamma\) | Plug-In | DiD | DiM | Plug-In | DiD | DiM | Plug-In | DiD | DiM | Plug-In | DiD | DiM |
| 1,000 | 1,000 | .99 | .0 | .00 | .00 | -.00 | .01 | .01 | .04 | .95 | .95 | .95 | .04 | .05 | .05 |
| 1,000 | 1,000 | .99 | .5 | .00 | .00 | -.00 | .01 | .01 | .06 | .95 | .95 | .95 | .04 | .06 | .05 |
| 1,000 | 1,000 | .50 | .0 | .00 | .00 | .00 | .04 | .04 | .05 | .94 | .95 | .94 | .06 | .05 | .05 |
| 1,000 | 1,000 | .50 | .5 | .00 | .00 | .00 | .05 | .05 | .06 | .95 | .95 | .95 | .06 | .05 | .05 |
| 1,000 | 1,000 | .00 | .0 | -.00 | .00 | -.00 | .04 | .07 | .04 | .95 | .94 | .95 | .05 | .06 | .05 |
| 1,000 | 1,000 | .00 | .5 | -.00 | .00 | -.00 | .06 | .07 | .06 | .95 | .95 | .95 | .04 | .05 | .05 |
| 25 | 25 | .99 | .0 | .00 | .00 | -.03 | .04 | .04 | .27 | .94 | .94 | .94 | .04 | .05 | .06 |
| 25 | 25 | .99 | .5 | .00 | -.01 | -.04 | .05 | .08 | .34 | .92 | .93 | .93 | .06 | .06 | .06 |
| 25 | 25 | .50 | .0 | -.01 | -.02 | -.02 | .24 | .29 | .26 | .94 | .95 | .94 | .04 | .04 | .05 |
| 25 | 25 | .50 | .5 | -.01 | -.01 | -.03 | .30 | .32 | .33 | .94 | .95 | .94 | .04 | .04 | .05 |
| 25 | 25 | .00 | .0 | -.03 | -.02 | -.03 | .28 | .38 | .27 | .95 | .95 | .93 | .06 | .04 | .06 |
| 25 | 25 | .00 | .5 | -.04 | -.02 | -.04 | .35 | .42 | .34 | .93 | .94 | .94 | .06 | .05 | .06 |
many other specifications: the standard deviation of DiD can be as much as 1.7 times larger than that of the plug-in efficient estimator, and the standard deviation of DiM can be as much as 7 times larger. These simulations thus illustrate how the plug-in efficient estimator can improve on DiD or DiM in cases where they are suboptimal, while retaining nearly identical performance when the DiD or DiM estimator is optimal.

### B. Simulations Based on Wood et al. (2020)

To evaluate the performance of our proposed methods in a more realistic staggered setting, we conduct simulations calibrated to our application in section IV, which is based on data from Wood et al. (2020). The outcome of interest $Y_{it}$ is the number of complaints against police officer $i$ in month $t$ for police officers in Chicago. Police officers were randomly assigned to first receive a procedural-justice training in period $G_i$. See section IV for more background on the application.

#### 1. Simulation Specification

We calibrate our baseline specification as follows. The numbers of observations and time periods in the data exactly match those used in our application. We set the untreated potential outcomes $Y_{it}(\infty)$ to match the observed outcomes in the data, $Y_{it}$ (which would exactly match the true potential outcomes if there were no treatment effect on any units). In our baseline simulation specification, there is no causal effect of treatment, so that $Y_{it}(g) = Y_{it}(\infty)$ for all $g$. (We describe an alternative simulation design with heterogeneous treatment effects in app. B.) In each
In total, there are 72 months of data on 5,537 officers. There are 47 distinct values of \( g \), with the cohort size \( N_g \) ranging from 3 to 575. In an alternative specification, we collapse the data to the yearly level, so that there are six time periods and five larger cohorts.

For each simulated dataset, we calculate the plug-in efficient estimator \( \hat{\theta}_S^{\text{CS}} \) for four estimands: the simple-weighted ATE (\( \theta_\text{simple} \)), the calendar- and cohort-weighted ATEs (\( \theta_\text{calendar} \) and \( \theta_\text{cohort} \)), and the instantaneous event-study parameter (\( \theta_\text{ES}^{0} \)).

### TABLE 3

| Estimator, Estimand | Bias | Coverage | FRT Size | Mean SE | SD |
|---------------------|------|----------|----------|---------|----|
| **Plug-in:**        |      |          |          |         |    |
| Calendar            | .01  | .93      | .07      | .26     | .28 |
| Cohort              | .00  | .92      | .06      | .26     | .28 |
| ES0                 | .00  | .96      | .04      | .32     | .31 |
| Simple              | .00  | .93      | .05      | .24     | .25 |
| **CS:**             |      |          |          |         |    |
| Calendar            | .01  | .95      | .06      | .51     | .52 |
| Cohort              | .02  | .95      | .04      | .47     | .46 |
| ES0                 | .00  | .96      | .04      | .44     | .43 |
| Simple              | .02  | .96      | .04      | .47     | .46 |
| **SA:**             |      |          |          |         |    |
| Calendar            | .00  | .91      | .04      | 1.44    | 1.50 |
| Cohort              | .01  | .90      | .05      | 1.51    | 1.58 |
| ES0                 | .00  | .96      | .04      | .91     | .94 |
| Simple              | .02  | .90      | .05      | 1.64    | 1.72 |

**Note.**—This table shows results for the plug-in efficient and CS and SA estimators in simulations calibrated to Wood et al. (2020). The estimands considered are the calendar-, cohort-, and simple-weighted ATEs, as well as the instantaneous event-study effect (ES0). Coverage refers to the fraction of the time a nominal 95% CI includes the true parameter, and FRT size refers to the null rejection rate of an FRT. Mean SE refers to the average estimated standard error, and SD refers to the actual standard deviation of the estimator. The bias, mean SE, and SD are all multiplied by 100 for ease of readability.

The CS estimator for ES0 corresponds with the estimator in de Chaisemartin and D’Haultfoeuille (2020).

Simulation draw \( s \), we randomly draw a vector of treatment dates \( G_s = (G_{s1}, \ldots, G_{sN}) \) such that the number of units first treated in period \( g \) matches that observed in the data (i.e., \( \sum_i G_i^g = N_g \) for all \( g \)). In total, there are 72 months of data on 5,537 officers. There are 47 distinct values of \( g \) with the cohort size \( N_g \) ranging from 3 to 575. In an alternative specification, we collapse the data to the yearly level, so that there are six time periods and five larger cohorts.

For each simulated dataset, we calculate the plug-in efficient estimator \( \hat{\theta}_S^{\text{CS}} \) for four estimands: the simple-weighted ATE (\( \theta_\text{simple} \)), the calendar- and cohort-weighted ATEs (\( \theta_\text{calendar} \) and \( \theta_\text{cohort} \)), and the instantaneous event-study parameter (\( \theta_\text{ES}^{0} \)).

(See sec. II.B for the formal definition of these estimands.) In our baseline specification, we use as \( \hat{X} \) the scalar weighted combination of pretreatment differences used by the Callaway and Sant’Anna (2021) estimator using not-yet-treated units as the comparison (\( \hat{X}_{\text{CS}}^{\text{S2}} \) in example 2). In the appendix, we also present results for...
an alternative specification in which $\hat{X}$ is a vector containing $\hat{\tau}_{r,g'}$ for all pairs $g, g' > t$. For comparison, we also compute the Callaway and Sant’Anna (2021) and Sun and Abraham (2021) estimators (CS and SA estimators, respectively) for the same estimand. Recall that for $v_{ES}$, the CS estimator coincides with the estimator proposed in de Chaisemartin and D’Haultfoeuille (2020) in our setting, since treatment is an absorbing state. The CIs are calculated as $\hat{v} \pm 1.96\hat{\sigma} / \sqrt{N}$ for the plug-in efficient estimator and analogously for the CS and SA estimators.25

2. Baseline Simulation Results

The results for our baseline specification are shown in tables 3 and 4. As seen in table 3, the plug-in efficient estimator is approximately unbiased, and 95% CIs based on our standard errors have coverage rates close to the nominal level for all of the estimands, with size distortions no larger than 3% for all of our specifications. The size for the FRT is also close to the nominal level, which is intuitive, since our baseline specification imposes the sharp null hypothesis, and thus the FRT should be exact up to simulation error. The CS and SA estimators are also both approximately unbiased and have coverage close to the nominal level, although coverage for the SA estimator is as low as 90% in some specifications.

Table 4 shows that there are large efficiency gains from using the plug-in efficient estimator, relative to using the CS or SA estimators. The table compares the standard deviation of the plug-in efficient estimator to those of the CS and SA estimators. Remarkably, using the plug-in efficient estimator reduces the standard deviation relative to the CS estimator by a

| Estimand   | Ratio of SD to That for Plug-In |
|------------|-------------------------------|
| Calendar   | 1.84 5.31                     |
| Cohort     | 1.67 5.72                     |
| ES0        | 1.39 3.02                     |
| Simple     | 1.85 6.86                     |

Note.—This table shows the ratio of the standard deviations of the CS and SA estimators to those of the plug-in efficient estimator, based on the simulation results in table 3.

25 The variance estimator for the CS and SA estimators is adapted analogously to that for the DiD and DiM estimators, as discussed in n. 23. We note that these design-based standard errors differ slightly from those proposed in the original papers (Callaway and Sant’Anna 2021; Sun and Abraham 2021), which adopt a sampling-based framework; using design-based standard errors makes the CIs for these estimators more directly comparable to those for the plug-in efficient estimator.
factor between 1.39 and 1.85, depending on the estimand. Since standard errors are proportional to the square root of the sample size for a fixed estimator, a reduction in standard errors by a factor of 1.85 roughly corresponds with an increase in sample size by a factor of 3.4. The gains of using the plug-in efficient estimator relative to the SA estimator are even larger, with reductions in the standard deviation by a factor of 3 or more. The reason for this is that the SA estimator uses only the last-treated units (rather than not-yet-treated units) as a comparison, but in our setting less than 1% of units are treated in the final period, leading to an efficiency loss.

3. Alternative Choices of $\hat{X}$

In appendix B, we present results where $\hat{X}$ is set to be a vector containing all possible comparisons of cohorts in periods before treatment. The dimension of this $\hat{X}$ is large relative to $N$ using monthly data, and in line with the discussion in remark 5, we find that the estimator has large bias and undercoverage owing to an overfitting problem. When the data are collapsed to the yearly level, the dimension of the $\hat{X}$ is more moderate, and the estimator is approximately unbiased and has good coverage, in line with the heuristic from Lei and Ding (2021) that the dimension of $\hat{X}$ should be small relative to $N^{-(1/2)}$. In our Monte Carlo simulation, however, the augmented $\hat{X}$ offers very minor precision gains relative to the $\hat{X}$ based on the CS estimator (Callaway and Sant’Anna 2021) used in our baseline specification. We thus focus on the latter choice of $\hat{X}$ in our application below.

4. Other Extensions

Appendix B contains several extensions to the baseline simulation specification, such as incorporating heterogeneous effects, annualizing the monthly data, and considering the other two outcomes in our application. As in the baseline specification, the plug-in efficient estimator has good coverage and offers efficiency gains relative to the other methods in nearly all specifications.

IV. Application to Procedural-Justice Training

A. Background

Reducing police misconduct and use of force is an important policy objective. Wood, Tyler, and Papachristos (2020) studied the Chicago Police

26 We also experimented with a 10-dimensional vector $\hat{X}$ that included our baseline scalar choice of $\hat{X}$ as the first element, as well as analogues to the baseline choice lagged by 1, ..., 9 periods, with very similar results to the baseline specification.
Department’s staggered rollout of a procedural-justice training program, which taught police officers strategies for emphasizing respect, neutrality, and transparency in the exercise of authority. Officers were randomly assigned a date for training. Wood, Tyler, and Papachristos (2020) found large and statistically significant impacts of the program on complaints and sustained complaints against police officers and on officer use of force. However, our reanalysis in Wood et al. (2020) highlighted a statistical error in the original analysis of Wood, Tyler, and Papachristos (2020), who failed to normalize for the fact that groups of officers trained in different months were of varying sizes. In Wood et al. (2020), we reanalyzed the data, using the procedure proposed by Callaway and Sant’Anna (2021) to correct for the error. The reanalysis found no significant effect on complaints or sustained complaints and borderline significant effects on use of force, although the CIs for all three outcomes included both near-zero and meaningfully large effects. Owens et al. (2018) studied a small pilot study of a procedural-justice training program in Seattle, with point estimates suggesting reductions in complaints but imprecisely estimated.

B. Data

We use the same data as in the reanalysis in Wood et al. (2020), which extends the data used in the original analysis of Wood, Tyler, and Papachristos (2020) through December 2016. As in Wood et al. (2020), we restrict attention to the balanced panel of officers who remained in the police force throughout the study period. We further drop officers in the initial pilot program and in special units, as these officers were trained in large batches and did not follow the random assignment protocol (see the supplementary material to Wood, Tyler, and Papachristos 2020). This leaves us a final sample of 5,537 officers.27 The data contain three outcome measures (complaints, sustained complaints, and use of force) at a monthly level for 72 months (6 years), with the first cohort trained in month 17 and the final cohort trained in the last month of the sample.

C. Estimation

We apply our proposed plug-in efficient estimator to estimate the effects of the procedural-justice training program on the three outcomes of interest. As in our Monte Carlo study, we use the scalar $\hat{X}$ such that $\beta = 1$ is

27 In the earlier working-paper version of this paper (Roth and Sant’Anna 2021), we included officers in the pilot program and special units, with qualitatively similar results. However, one can formally reject the null hypothesis of random assignment when including these officers (see table 6).
the CS estimator \((\hat{\tau}^{CS}; \text{Callaway and Sant’Anna 2021})\). We estimate the simple, cohort, and calendar-weighted average effects described in section II.B and used in our Monte Carlo study. We also estimate the event-study effects for the first 24 months after treatment, which includes the instantaneous event-study effect studied in our Monte Carlo as a special case (for event-time 0). For comparison, we also estimate the CS estimator as in Wood et al. (2020).\(^{28}\)

D. Results

1. Baseline Results

Figure 1 shows the results of our analysis for the three aggregate summary parameters. Table 5 compares the magnitudes of these estimates and their 95% CIs to the mean of the outcome in the 12 months before the pilot program began. It also reports \(p\)-values from the FRT.

For all outcomes, the CIs for the plug-in efficient estimator overlap with those of the CS estimator (Callaway and Sant’Anna 2021) but are substantially narrower. Indeed, the final column of table 5 shows that the standard errors (or equivalently, the length of the CIs) range from 1.4–8.4 times smaller than those of the CS estimator, depending on the specification. As in Wood et al. (2020), we find no significant impact on complaints using any of the aggregations. Our bounds on the magnitude of the treatment effect are substantially tighter than before, however. For instance, using the simple aggregation we can now rule out reductions in complaints of more than 13%, compared with a bound of 33% using the CS estimator, and our standard errors are roughly half as large as when using the CS estimator. For use of force, the point estimates from the efficient estimator are somewhat smaller (in magnitude) than those from the CS estimator, but they suggest a reduction in use of force of around 16%–18% of the pretreatment mean. However, the upper bounds of the CIs are close to zero; \(p\)-values using the FRT are between 0.02 and 0.06. Thus, although precision is substantially higher than when using the CS estimator, the CIs for use of force still include effects from near zero up to about 30% of the pretreatment mean. For sustained complaints, all of the point estimates are near zero and the CIs are substantially narrower than when using the CS estimator, although the plug-in efficient estimate using the calendar aggregation is marginally significant (FRT \(p\)-value = .04). In figures C.1 and C.2, we show event-study plots using the plug-in efficient and CS estimates. The figures do

\(^{28}\) The CS estimates are not identical to those in Wood et al. (2020) for two reasons (although they are qualitatively similar). The first is that we exclude officers in the pilot program and special units. Second, for direct comparability, we calculate design-based standard errors for the CS estimator using the analogue to \(\hat{\sigma}_{se}\), and thus the reported SEs differ slightly from the sampling-based SEs reported in Wood et al. (2020).
not show a clear significant effect for any of the outcomes, nor do they show significant placebo pretreatment effects.

2. Balance and Robustness Checks

Although treatment timing was explicitly randomized in our application, as discussed in the supplement to Wood, Tyler, and Papachristos (2020), there are some concerns about noncompliance wherein officers could volunteer to receive the training before their randomly assigned date, particularly toward the end of the training period. (The observed treatment variable in the data is the actual training date, and whether an officer volunteered is not recorded.) We therefore conduct a series of robustness and balance checks to evaluate the extent to which noncompliance may have violated the assumption of random treatment timing. We first test for balance in pretreatment outcomes by testing the null that \( \mathbb{E} [\hat{X}] = 0 \), as described in section II.H. In particular, we use the scalar \( \hat{X} \) used by the CS estimator for each of our summary parameters and outcomes, with results shown in table 6. Reassuringly, we do not find any (individual or jointly) significant imbalances in \( \hat{X} \) using our main analysis sample. Interestingly, we do find a significant imbalance for use of force.
### Table 5
Estimates and 95% CIs as a Percentage of Pretreatment Means

| Outcome, Estimand | Pretreatment Mean | Plug-In | CS |
|-------------------|-------------------|---------|-----|
|                   | Estimate (%) | LB (%) | UB (%) | Estimate (%) | LB (%) | UB (%) | Estimate (%) | LB (%) | UB (%) |
| Complaints:       |               |         |       |               |         |       |               |         |       |
| Simple            | .052          | -5      | -13   | 4              | .326    | -16   | -33   | 1              | .063    | 2.0    |
| Calendar          | .052          | -1      | -15   | 13             | .882    | -12   | -35   | 11             | .299    | 1.6    |
| Cohort            | .052          | -3      | -13   | 6              | .485    | -16   | -33   | 1              | .059    | 1.8    |
| Sustained:        |               |         |       |               |         |       |       |               |         |       |
| Simple            | .004          | -15     | -33   | 2              | .102    | 2     | -75   | 79             | .955    | 4.3    |
| Calendar          | .004          | -17     | -34   | -0             | .038    | -52   | -194  | 90             | .493    | 8.4    |
| Cohort            | .004          | -13     | -31   | 5              | .170    | 3     | -71   | 76             | .935    | 4.2    |
| Use of Force:     |               |         |       |               |         |       |       |               |         |       |
| Simple            | .051          | -18     | -32   | -3             | .031    | -25   | -46   | -4             | .021    | 1.5    |
| Calendar          | .051          | -16     | -30   | -2             | .020    | -30   | -55   | -4             | .022    | 1.8    |
| Cohort            | .051          | -16     | -31   | -1             | .063    | -24   | -46   | -3             | .024    | 1.4    |

Note.—This table shows the pretreatment means for the three outcomes. It also displays the estimates and 95% CIs in fig. 1 as percentages of these means, as well as the $p$-value from an FRT. The final column shows the ratio of the length of the CI for the CS estimator relative to that for the plug-in efficient estimator. All FRT $p$-values are based on 5,000 permutations. LB = lower bound; UB = upper bound.
| OUTCOME, ESTIMAND | MAIN ESTIMATION SAMPLE |            |            |            |            | INCLUDING PILOT AND SPECIAL |            |            |            |
|-------------------|------------------------|------------|------------|------------|------------|-----------------------------|------------|------------|------------|
|                   |                        | Û           | t-STATISTIC| p-VALUE    | p-VALUE (FRT) |                | Û           | t-STATISTIC| p-VALUE    | p-VALUE (FRT) |                |
| Complaints:       |                        | .007       | 1.55       | .122       | .125        | .162           | .005       | 1.22       | .221       | .223         | .441          |
| Simple            |                        | .008       | 1.76       | .078       | .079        | .162           | .004       | 1.04       | .299       | .298         | .441          |
| Cohort            |                        | .006       | 1.22       | .222       | .227        | .162           | .010       | 1.30       | .192       | .243         | .441          |
| Calendar          |                        | .004       | 1.27       | .205       | .207        | .162           | .003       | 1.03       | .304       | .298         | .441          |
| Sustained:        |                        |            |            |            |            |                |            |            |            |              |               |
| Simple            |                        | -.001      | .46        | .645       | .653        | .883           | -.001      | .97        | .331       | .333         | .541          |
| Cohort            |                        | -.001      | .43        | .666       | .678        | .883           | -.002      | 1.03       | .302       | .303         | .541          |
| Calendar          |                        | .002       | .48        | .629       | .665        | .883           | -.001      | .42        | .678       | .722         | .541          |
| ES0               |                        | .000       | .23        | .816       | .810        | .883           | .000       | .32        | .751       | .746         | .541          |
| Use of force:     |                        |            |            |            |            |                |            |            |            |              |               |
| Simple            |                        | .005       | .91        | .361       | .359        | .360           | .005       | 1.04       | .300       | .309         | .017          |
| Cohort            |                        | .006       | 1.10       | .272       | .271        | .360           | .004       | .91        | .364       | .371         | .017          |
| Calendar          |                        | .008       | 1.22       | .223       | .222        | .360           | .013       | 1.59       | .112       | .149         | .017          |
| ES0               |                        | .005       | 1.28       | .201       | .200        | .360           | .008       | 2.91       | .004       | .002         | .017          |

**Note.**—This table shows balance on pretreatment outcomes by testing the null hypothesis that \( E[\tilde{X}] = 0 \). The columns report the value of \( \tilde{X} \), its \( t \)-statistic, the \( p \)-value based on the \( t \)-statistic, the \( p \)-value using an FRT, and a \( p \)-value for the joint test that \( E[\tilde{X}] = 0 \) for all estimands using the same outcome (computed using an FRT with the \( \max |t| \) statistic). The columns labeled “Main Estimation Sample” use the main data for our analysis, whereas those labeled “Including Pilot and Special” include officers in the pilot program and special units, who did not follow the randomization protocol. All FRT \( p \)-values are based on 5,000 permutations.
if we include officers in the pilot program and special units, who are known not to have followed the randomization protocol, which suggests that these tests may be powered to detect some relevant violations of the randomization assumption. Second, we construct an “event-study plot” that tests for placebo pretreatment effects before the date of training, and we generally do not find any concerning pretreatment placebo effects. These results, and those for our subsequent balance checks, are shown in appendix C. Third, we test for covariate balance on year of birth, one of the few pretreatment demographic variables in the data. We find that average year of birth is similar across training dates, although in one of our two specifications we statistically reject the null of exact equality at the 10% level ($p$-value = .08), possibly suggesting some slight imbalance on age. Finally, as a robustness check, we redo our main analysis excluding units who were trained in the final year of the training program, when noncompliance was suspected to be more severe. The qualitative patterns are similar, although the estimates for use of force are no longer statistically significant in some specifications.

3. Implications

Our analysis provides the most precise estimates to date on the effectiveness of procedural-justice training for police officers. Our estimates for the effects of the program on complaints against officers are close to zero, with much tighter upper bounds on the effectiveness at reducing complaints than in previous work. The results for use of force are more mixed, with point estimates suggesting reductions of 16%–18% but with CIs that include near-zero or zero effects in all specifications. Thus, more research is needed to determine whether procedural-justice training can be a useful tool in meaningfully reducing officer use of force. We encourage police departments planning to implement such training in the future to consider a randomized staggered rollout, which is a potentially low-cost way to learn more about the effectiveness of the program.

V. Conclusion

This paper considers efficient estimation in settings with staggered adoption and (quasi-)random treatment timing. The assumption of (quasi-)random treatment timing is technically stronger than the parallel-trends assumption, but it is often the justification given for the parallel-trends assumption in practice, and it can be ensured by design in experimental contexts where the researcher controls the timing of treatment. We derive the most efficient estimator in a large class of estimators that nests many existing approaches. The “oracle” efficient estimator is not known in practice, but we show that a plug-in sample analogue has similar properties in
large populations, and we derive both $t$-based and permutation-based approaches to inference. We find in simulations that the proposed plug-in efficient estimator is approximately unbiased, yields reliable inference, and substantially increases precision relative to existing methods. We apply our proposed methodology to obtain the most precise estimates to date of the causal effects of procedural-justice training programs for police officers.

Data Availability

Code and data replicating the tables and figures in this article can be found in Roth and Sant’Anna (2023a) in the Harvard Dataverse, https://doi.org/10.7910/DVN/ICESNW.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. “Sampling-Based versus Design-Based Uncertainty in Regression Analysis.” *Econometrica* 88 (1): 265–96.

Abbring, Jaap H., and Gerard J. van den Berg. 2003. “The Nonparametric Identification of Treatment Effects in Duration Models.” *Econometrica* 71 (5): 1491–517.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton, NJ: Princeton Univ. Press.

Athey, Susan, and Guido W. Imbens. 2022. “Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption.” *J. Econometrics* 226 (1): 62–79.

Bai, Yuehao, Joseph P. Romano, and Azeem M. Shaikh. 2022. “Inference in Experiments with Matched Pairs.” *J. American Statis. Assoc.* 117 (540): 1726–37.

Borusyak, Kirill, and Peter Hull. 2020. “Non-random Exposure to Exogenous Shocks: Theory and Applications.” Working Paper no. 27845 (September), NBER, Cambridge, MA.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. Forthcoming. “Revisiting Event Study Designs Robust and Efficient Estimation.” *Rev. Econ. Studies*.

Brown, Celia A., and Richard J. Lilford. 2006. “The Stepped Wedge Trial Design: A Systematic Review.” *BMC Medical Research Methodology* 6:54.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2018. “Inference under Covariate-Adaptive Randomization.” *J. American Statis. Assoc.* 113 (524): 1784–96.

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

Chung, EunYi, and Joseph P. Romano. 2013. “Exact and Asymptotically Robust Permutation Tests.” *Ann. Statist.* 41 (2): 484–507.

———. 2016. “Multivariate and Multiple Permutation Tests.” *J. Econometrics* 193 (1): 76–91.
de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *A.E.R.* 110 (9): 2964–96.

——. 2021. “Difference-in-Differences Estimators of Intertemporal Treatment Effects.” Working paper (May). https://arxiv.org/abs/2007.04267v6.

Deshpande, Manasi, and Yue Li. 2019. “Who Is Screened Out? Application Costs and the Targeting of Disability Programs.” *American Econ. J. Econ. Policy* 11 (4): 213–48.

DiCiccio, Cyrus J., and Joseph P. Romano. 2017. “Robust Permutation Tests for Correlation and Regression Coefficients.” *J. American Statis. Assoc.* 112 (519): 1211–20.

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

Doleac, Jennifer. 2020. “How to Fix Policing.” Commentary, Niskanen Center, Washington, DC. https://www.niskanencenter.org/how-to-fix-policing/.

Fadlon, Itzik, and Torben Heien Nielsen. 2021. “Family Labor Supply Responses to Severe Health Shocks: Evidence from Danish Administrative Records.” *American Econ. J. Appl. Econ.* 13 (3): 1–50.

Freedman, David A. 2008a. “On Regression Adjustments in Experiments with Several Treatments.” *Ann. Appl. Statis.* 2 (1): 176–96.

——. 2008b. “On Regression Adjustments to Experimental Data.” *Advances Appl. Math.* 40 (2): 180–93.

Frison, Lars, and Stuart J. Pocock. 1992. “Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design.” *Statis. and Probability Letters* 11 (13): 1685–704.

Funatogawa, Takashi, Ikuko Funatogawa, and Yu Shyr. 2011. “Analysis of Covariance with Pre-treatment Measurements in Randomized Trials under the Cases That Covariances and Post-treatment Variances Differ between Groups.” *Biometrical J.* 53 (3): 512–24.

Glynn, Adam N., and Kevin M. Quinn. 2010. “An Introduction to the Augmented Inverse Propensity Weighted Estimator.” *Polit. Analysis* 18 (1): 36–56.

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

Hahn, Jinyong. 1998. “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects.” *Econometrica* 66 (2): 315–31.

Imbens, Guido W. 2004. “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review.” *Rev. Econ. and Statis.* 86 (1): 4–29.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge Univ. Press.

Janssen, Arnold. 1997. “Studentized Permutation Tests for Non-i.i.d. Hypotheses and the Generalized Behrens-Fisher Problem.” *Statis. and Probability Letters* 36 (1): 9–21.

Ji, Xinyao, Gunther Fink, Paul Jacob Robyn, and Dylan S. Small. 2017. “Randomization Inference for Stepped-Wedge Cluster-Randomized Trials: An Application to Community-Based Health Insurance.” *Ann. Appl. Statis.* 11 (1): 1–20.

Lechner, Michael. 2010. “The Estimation of Causal Effects by Difference-in-Difference Methods.” *Found. and Trends Econometrics* 4 (3): 165–224.

Lei, Lihua, and Peng Ding. 2021. “Regression Adjustment in Completely Randomized Experiments with a Diverging Number of Covariates.” *Biometrika* 108 (4): 815–28.

Li, Xinran, and Peng Ding. 2017. “General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference.” *J. American Statis. Assoc.* 112 (520): 1759–69.
Lin, Winston. 2013. “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique.” Ann. Appl. Statis. 7 (1): 295–318.

Lindner, Stephan, and K. John McConnell. 2021. “Heterogeneous Treatment Effects and Bias in the Analysis of the Stepped Wedge Design.” Health Services and Outcomes Res. Methodology 21:419–38.

MacKinnon, James G., and Matthew D. Webb. 2020. “Randomization Inference for Difference-in-Differences with Few Treated Clusters.” J. Econometrics 218 (2): 435–50.

Malani, Anup, and Julian Reif. 2015. “Interpreting Pre-trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform.” J. Public Econ. 124:1–17.

Manski, Charles F., and John V. Pepper. 2018. “How Do Right-To-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions.” Rev. Econ. and Statis. 100 (2): 232–44.

McKenzie, David. 2012. “Beyond Baseline and Follow-Up: The Case for More T in Experiments.” J. Development Econ. 99 (2): 210–21.

Negi, Akanksha, and Jeffrey M. Wooldridge. 2021. “Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects.” Econometric Rev. 40 (5): 504–34.

Nekoei, Arash, and David Seim. 2023. “How Do Inheritances Shape Wealth Inequality? Theory and Evidence from Sweden.” Rev. Econ. Studies 90 (1): 463–98.

Neyman, Jerzy. 1990 (1923). “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” Translated and edited by D. M. Dabrowska and T. P. Speed. Statis. Sci. 5 (4): 465–72.

Owens, Emily, David Weisburd, Karen L. Amendola, and Geoffrey P. Alpert. 2018. “Can You Build a Better Cop? Experimental Evidence on Supervision, Training, and Policing in the Community.” Criminology and Public Policy 17 (1): 41–87.

Parker, Jonathan A., Nicholas S. Souleles, David S. Johnson, and Robert McClelland. 2013. “Consumer Spending and the Economic Stimulus Payments of 2008.” A.E.R. 103 (6): 2530–53.

Rambachan, Ashesh, and Jonathan Roth. 2020. “Design-Based Uncertainty for Quasi-Experiments.” Working paper (August). https://arxiv.org/abs/2008.00602v2.

Roth, Jonathan. 2022. “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends.” A.E.R. Insights 4 (3): 305–22.

Roth, Jonathan, and Pedro H. C. Sant’Anna. 2021. “Efficient Estimation for Staggered Rollout Designs.” Working paper (June). https://arxiv.org/abs/2102.01291v3.

———. 2023a. “Replication Data for: ‘Efficient Estimation for Staggered Rollout Designs.’” Harvard Dataverse, https://doi.org/10.7910/DVN/ICESNW.

———. 2023b. “When Is Parallel Trends Sensitive to Functional Form?” Econometrica 91 (2): 737–47.

Sant’Anna, Pedro H. C., and Jun Zhao. 2020. “Doubly Robust Difference-in-Differences Estimators.” J. Econometrics 219 (1): 101–22.

Sekhon, Jasjeet S., and Yotam Shem-Tov. 2021. “Inference on a New Class of Sample Average Treatment Effects.” J. American Statis. Assoc. 116 (534): 798–804.

Shaikh, Azeem M., and Panos Toulis. 2021. “Randomization Tests in Observational Studies with Staggered Adoption of Treatment.” J. American Statis. Assoc. 116 (536): 1835–48.

Sun, Liyang, and Sarah Abraham. 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” J. Econometrics 225 (2): 175–99.
Wan, Fei. 2020. “Analyzing Pre-Post Designs Using the Analysis of Covariance Models with and without the Interaction Term in a Heterogeneous Study Population.” *Statis. Methods Medical Res.* 29 (1): 189–204.

Wood, George, Tom R. Tyler, and Andrew V. Papachristos. 2020. “Procedural Justice Training Reduces Police Use of Force and Complaints against Officers.” *Proc. Nat. Acad. Sci. USA* 117 (18): 9815–21.

Wood, George, Tom R. Tyler, Andrew V. Papachristos, Jonathan Roth, and Pedro H. Sant’Anna. 2020. “Revised Findings for ‘Procedural Justice Training Reduces Police Use of Force and Complaints against Officers.’” https://www.jonathandroth.com/assets/files/wood-etal-revisited.pdf.

Wu, Jason, and Peng Ding. 2021. “Randomization Tests for Weak Null Hypotheses in Randomized Experiments.” *J. American Statis. Assoc.* 116 (536): 1898–913.

Xiong, Ruoxuan, Susan Athey, Mohsen Bayati, and Guido Imbens. 2019. “Optimal Experimental Design for Staggered Rollouts.” Working paper (November). https://arxiv.org/abs/1911.03764v1.

Yang, Li, and Anastasios A. Tsiatis. 2001. “Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial.” *American Statistician* 55 (4): 314–21.

Zhao, Anqi, and Peng Ding. 2021. “Covariate-Adjusted Fisher Randomization Tests for the Average Treatment Effect.” *J. Econometrics* 225 (2): 278–94.