Efficient Estimation for Staggered Rollout Designs*

Jonathan Roth†   Pedro H.C. Sant'Anna‡

February 3, 2021

Abstract

Researchers are often interested in the causal effect of treatments that are rolled out to different units at different points in time. This paper studies how to efficiently estimate a variety of causal parameters in such staggered rollout designs when treatment timing is (as-if) randomly assigned. We solve for the most efficient estimator in a class of estimators that nests two-way fixed effects models as well as several popular generalized difference-in-differences methods. The efficient estimator is not feasible in practice because it requires knowledge of the optimal weights to be placed on pre-treatment outcomes. However, the optimal weights can be estimated from the data, and in large datasets the plug-in estimator that uses the estimated weights has similar properties to the "oracle" efficient estimator. We illustrate the performance of the plug-in efficient estimator in simulations and in an application to Wood, Tyler and Papachristos (2020a)'s study of the staggered rollout of a procedural justice training program for police officers. We find that confidence intervals based on the plug-in efficient estimator have good coverage and can be as much as five times shorter than confidence intervals based on existing methods. As an empirical contribution of independent interest, our application provides the most precise estimates to date on the effectiveness of procedural justice training programs for police officers.

---

†Microsoft. Jonathan.Roth@microsoft.com

‡Vanderbilt University. pedro.h.santanna@vanderbilt.edu

1 Introduction

Researchers are often interested in the causal effects of a treatment that has a staggered rollout, meaning that it is first implemented for different units at different times. For instance, social scientists may be interested in the causal effect of a policy that is adopted in different states at different times. Businesses may likewise be interested in the causal effect of a new feature or advertising campaign that is introduced to different customers over time. In many cases, the timing of the rollout is controlled by the researcher and can be explicitly randomized. In others, researchers argue that the timing of the treatment is as-if randomly assigned.

In these settings, researchers often estimate treatment effects using methods that extend the simple two-period difference-in-differences estimator to the staggered setting. It is common practice to estimate causal effects using two-way fixed effects (TWFE) models that control for both time and unit fixed effects (e.g. Xiong, Athey, Bayati and Imbens, 2019). Recent work has shown, however, that the estimand of TWFE models may be difficult to interpret under treatment effect heterogeneity (see Related Literature below). The literature has therefore proposed a variety of alternative procedures that yield more easily-interpretable estimands under heterogeneous treatment effects (Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020). All of these procedures exploit a generalized "parallel trends" assumption for estimation. However, the assumption of random treatment timing is stronger than parallel trends. This suggests that it might be possible to obtain more precise estimates by more fully exploiting the random timing of treatment.

This paper considers efficient estimation of causal effects in settings where the timing of treatment is (as-if) randomly assigned. We begin by introducing a design-based framework that formalizes the notion that treatment timing is (as-if) randomly assigned. We consider estimation of a variety of causal estimands using a class of estimators that nests canonical two-way fixed effects models as well as the alternative estimators discussed above as special cases. We then solve for the most efficient estimator in this class. The efficient estimator more fully exploits the implications of random treatment timing, which is stronger than the generalized parallel trends assumption on which conventional estimators are based. As a result, the efficient estimator we derive asymptotically dominates conventional estimation approaches for the same estimand in terms of efficiency, with large gains in Monte Carlo simulations. We therefore recommend use of the efficient estimator in settings where treatment timing is either random by design or assumed to be quasi-random.

For clarity of exposition and to connect our results to previous work, we begin by analyzing the canonical two-period difference-in-differences model under randomized treatment.

All units are untreated in the first period, and a subset of units are randomly assigned to begin treatment in the second period. We consider estimators of the average treatment effect (ATE) of the form $\hat{\theta}_\beta = (\bar{Y}_{11} - \bar{Y}_{01}) - \beta(\bar{Y}_{10} - \bar{Y}_{00})$, where $\bar{Y}_{dt}$ is the sample mean of the outcome for treatment group d in period t. These estimators take the simple difference in means in period 1 and then adjust linearly for the difference in means in period 0. The canonical difference-in-differences estimator corresponds with the special case of $\beta = 1$. Under the assumption that period 0 outcomes are unaffected by treatment status in period 1 (i.e. there are no anticipatory effects of treatment), the period 0 outcomes are isomorphic to fixed covariates in a random experiment. We can then apply results from Lin (2013) on covariate adjustment in random experiments to (i) show that $\hat{\theta}_\beta$ is unbiased for the ATE for all β, and (ii) solve for the variance-minimizing value β^*, which depends on the covariance between the (treated and untreated) potential outcomes in period 1 and the pre-treatment outcomes. In general, the efficient value β^* will not be equal to 1, and thus the DiD estimator will be inefficient. Although the "oracle" β^* will generally not be known, as in Lin (2013), a plug-in estimator based on a sample analog of β^* will achieve the efficient variance in large populations.

We next consider the more practically relevant case in which there is staggered timing across multiple periods. There are T periods, and unit i is first treated in period $G_i \in \mathcal{G} \subseteq \{1, ..., T, \infty\}$, with $G_i = \infty$ denoting that i is never treated. There are many possible ways of aggregating treatment effects across cohorts and time periods in the staggered treatment setting, and so we consider a broad class of estimands that encompasses many possible aggregation schemes. Specifically, we define $\tau_{t,gg'}$ to be the average effect on the outcome in period t of changing the initial treatment date from g' to g. We then consider the class of estimands that are linear combinations of these building blocks, $\theta = \sum_{t,g,g'} a_{t,g,g'} \tau_{t,gg'}$. Our framework thus accommodates a variety of summary measures of dynamic treatment effects, including several aggregation schemes proposed in the recent literature.

We consider the class of estimators that start with a sample analog to the target parameter and then adjust by a linear combination of pre-treatment outcomes. More precisely, we consider estimators of the form $\hat{\tau}_\beta = \sum_{t,g} a_{t,g,g'} \hat{\tau}_{t,gg'} - \hat{X}'\beta$, where the first term in $\hat{\tau}_\beta$ replaces the $\tau_{t,gg'}$ with their sample analogs in the definition of θ, and the second term adjusts for a linear combination of $\hat{X}$, where $\hat{X}$ is a vector that compares outcomes for cohorts treated at different dates at points in time before either was treated. We show that a variety of estimation procedures are part of this class for an appropriately defined $\hat{X}$, including the classical TWFE model as well as recent procedures proposed by Sun and Abraham (2020), de Chaisemartin and D'Haultfœuille (2020), and Callaway and Sant'Anna (2020). All estimators of this form are unbiased for θ under the assumptions of random treatment timing

and no anticipation.

We then derive the most efficient estimator in this class. The optimal coefficient β^* depends on covariances between the potential outcomes over time, and thus in general will not coincide with the fixed coefficients in any of the previously proposed procedures discussed above. As in the two-period case, the "oracle" value β^* will typically not be known ex ante, and will need to be replaced with a sample analog $\hat{\beta}^*$. Similar to the two-period case, we show that using the plug-in estimator is asymptotically unbiased and efficient under large-population asymptotics, exploiting the generalized finite-population central limit theorems in Li and Ding (2017). In a Monte Carlo study calibrated to our application, we find that confidence intervals based on the plug-in efficient estimator have good coverage properties and are substantially shorter than the procedures of Callaway and Sant'Anna (2020) and de Chaisemartin and D'Haultfœuille (2020).

As an illustration of our method and standalone empirical contribution, we re-examine the effectiveness of procedural justice training for police officers. We use data from Wood et al. (2020a), who studied the randomized rollout of a procedural justice training program in Chicago. The original study by Wood et al. (2020a) found that the program produced large and highly statistically significant reductions in (sustained) complaints against police officers and officer use of force. These findings have been influential in policy discussions about police reform (e.g. Doleac, 2020). However, an earlier version of our analysis revealed a statistical error in the analysis of Wood et al. (2020a), which did not account for the fact that cohorts trained on different days were of different sizes, leading to spuriously large estimates. In Wood, Tyler, Papachristos, Roth and Sant'Anna (2020b), we worked with the authors of the original paper to re-analyze the data using the state-of-the art tools in Callaway and Sant'Anna (2020). Our re-analysis found no significant effects on complaints or sustained complaints, and borderline significant effects on police use of force, although the confidence intervals for all three outcomes were large and included both economically small and meaningful effects. In this paper, we re-analyze the data using our proposed methodology.

We find that the use of our proposed methodology allows us to obtain substantially more precise estimates of the effect of the training program, leading to a reduction in standard errors by a factor between 1.3 and 5.6 depending on the specification. We again find limited evidence of a meaningful effect of the program on complaints or sustained complaints and borderline significant overall effects on use of force. Our revised estimates have much greater precision than in our previous analysis, however. For example, our baseline estimate for the overall average effect on complaints (using a simple aggregation across time periods and cohorts) is -2% relative to the pre-treatment mean with a 95% CI of [-11,6], compared with

an estimate of -10% and CI of [-26,5] using the procedure of Callaway and Sant'Anna (2020). Likewise, we find a borderline significant effect on use of force of -15% (CI: [-29,0]), compared with our previous estimate of -22% (CI: [-43,-2]). We caution, however, that the marginally significant results for use of force are not significant after adjusting for testing hypotheses on multiple outcomes.

Related Literature. This paper contributes to an active literature on difference-in-differences and related methods with staggered treatment timing. Several recent papers have illustrated that the estimand of standard TWFE models may not have an intuitive causal interpretation when there are heterogeneous treatment effects, and new estimators for more sensible causal estimands have been introduced (Athey and Imbens, 2018; Borusyak and Jaravel, 2016; Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2018; Imai and Kim, 2020; Meer and West, 2016; Słoczyński, 2020; Sun and Abraham, 2020). In contrast to most of the previous literature, we consider the efficiency of various procedures under random treatment timing. This assumption is stronger than the generalized parallel trends assumptions considered in previous work, and thus our proposed method will not be applicable in settings where the researcher is confident in parallel trends but not in random treatment timing.[1] On the other hand, under suitable regularity conditions our proposed plug-in efficient estimator is at least as efficient in large populations as the methods proposed in previous work when treatment timing is (as-if) randomly assigned, and will often be substantially more precise.

Additionally, most of the pre-existing literature has adopted a sampling-based perspective, where uncertainty in the data arises from the sample drawn from a superpopulation. By contrast, we adopt a design-based framework in which the population is fixed and uncertainty arises from the randomness of treatment timing. This framework is useful for formalizing the notion of random treatment timing, and may be especially appealing in settings where the superpopulation is not clear, such as when the researcher has access to all counties in the United States (Manski and Pepper, 2018). Athey and Imbens (2018) adopt a design-based framework similar to ours, but consider the interpretation of the estimand of two-way fixed effects models rather than efficient estimation. Shaikh and Toulis (2019) consider inference on sharp null hypotheses in a design-based model where treatment timing is random conditional on observables; by contrast, we consider inference on average causal effects under unconditional random treatment timing.

Several papers in both the economics and biostatistics literatures study the efficiency of

---

[1] In Roth and Sant'Anna (2021), we show that if treatment timing is not random, then the parallel trends assumption will be sensitive to functional form without strong assumptions on the full distribution of potential outcomes.

various procedures in non-staggered settings similar to that in Section 2. Frison and Pocock (1992) and McKenzie (2012) provide conditions in a sampling-based model under which the efficiency of difference-in-differences is dominated by that of a non-interacted Analysis of Covariance (ANCOVA) model, which regresses the post-treatment outcome on treatment status and lagged outcomes. Their procedure has the same asymptotic efficiency as our proposed procedure under homogenous treatment effects, but is generally less efficient if there is non-zero finite-population covariance between treatment effects and pre-treatment outcomes. Indeed, in the simple setting considered in Section 2, our proposed procedure is equivalent to an "interacted ANCOVA" which includes lagged outcomes interacted with treatment status. Wan (2020) shows that the efficiency of the interacted ANCOVA dominates that of the non-interacted ANCOVA in a sampling-based model that assumes normally distributed outcomes. Funatogawa, Funatogawa and Shyr (2011) give an informal derivation of a parameter analogous to the oracle slope β^* in a sampling-based model; however, they do not consider estimation or inference when the oracle is unknown. None of the aforementioned papers considers a design-based framework, nor do they study the common case of staggered treatment timing as we do in Section 3.

Our work is also related to Xiong et al. (2019) and Basse, Ding and Toulis (2020), who consider how to optimally design 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. Ding and Li (2019) show a bracketing relationship between the biases of difference-in-differences and other estimators in the class we consider when treatment timing is not random, but do not consider efficiency under random treatment timing.

Finally, we contribute to the literature on the effectiveness of procedural justice training programs for police officers. Previous work has studied the program in Chicago that we study (Wood et al., 2020a,b) and a smaller pilot evaluation in Seattle (Owens, Weisburd, Amendola and Alpert, 2018). Although qualitatively in line with previous findings in the literature, our analysis provides by far the most precise estimates from a randomized evaluation. For example, the standard error for our estimate for the effect on citizen complaints, measured as a percentage of the pre-treatment mean, is 1.9 times smaller than the estimate in Wood et al. (2020b) and over 3 times smaller than that in Owens et al. (2018).

2   Simple case: Two periods

We begin by developing intuition for our more general results by considering a canonical two-period difference-in-differences model. All of the results in this section can be viewed as special cases of the more general results for staggered rollouts in Section 3. We provide

proofs where we think they will aid in developing intuition, but defer some of the more technical proofs to the theorems in the following section.

2.1 Model

There is a finite population of N units. We observe data for 2 periods $t = 0, 1$. All units are untreated in $t = 0$, and some units receive a treatment of interest in $t = 1$. We denote by $Y_{it}(1), Y_{it}(0)$ the potential outcomes for unit i in period t under treatment and control, respectively, and we observe the outcome $Y_{it} = D_i Y_{it}(1) + (1 - D_i)Y_{it}(0)$, where D_i is an indicator for whether unit i is treated.[2] Following Neyman (1923) for randomized experiments and Athey and Imbens (2018) and Rambachan and Roth (2020) for DiD designs, we treat as fixed (or condition on) the potential outcomes and the number of treated and untreated units (N_0 and N_1); the only source of uncertainty in our model comes from the vector of treatment assignments $D = (D_1, ..., D_N)$, which is stochastic. All expectations ($\mathbb{E}[\cdot]$) and probability statements ($\mathbb{P}(\cdot)$) are taken over the distribution of D conditional on the number of treated units (N_1) and the potential outcomes, although we suppress this conditioning unless needed for clarity. For a non-stochastic 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 $\mathbb{V}\mathrm{ar}_f[W_i] = (N-1)^{-1}\sum_i(W_i - \mathbb{E}_f[W_i])(W_i - \mathbb{E}_f[W_i])'$ the finite-population expectation and variance of W_i.

The target parameter of interest is the average treatment effect in $t = 1$,

$$\tau_1 := \frac{1}{N}\sum_i (Y_{it=1}(1) - Y_{it=1}(0)).$$

We now introduce two assumptions that we will maintain throughout our analysis. We first assume that the assignment of treatment status is random.[3]

Assumption 1 (Random treatment assignment (2 periods)). $\mathbb{P}(D = d) = 1/\binom{N}{N_1}$ if $\sum_i d_i = N_1$, and 0 otherwise.

We also assume that treatment status has no effect on outcomes in $t = 0$, before treatment 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 (Malani and Reif, 2015).

Assumption 2 (No anticipation (2 periods)). *For all i, $Y_{it=0}(0) = Y_{it=0}(1)$.*

---

[2]In Section 3, we will index potential outcomes by the date of treatment timing, so $Y_{it}(0)$ in this section corresponds with $Y_{it}(\infty)$ in the notation of Section 3. We use the notation $Y_{it}(0)$ here to make the connections to the literature on randomized experiments more explicit.

[3]Note that we condition on the number of treated units (N_1), so in contrast to standard sampling-based approaches, unit i's treatment status D_i is correlated with that of unit j.

2.2 Efficient Estimation and Comparison to DiD

The canonical difference-in-differences estimator is

$$\hat{\tau}^{DiD} = (\bar{Y}_{t=1}^1 - \bar{Y}_{t=1}^0) - (\bar{Y}_{t=0}^1 - \bar{Y}_{t=0}^0), \tag{1}$$

where $\bar{Y}_t^1 = N_1^{-1} \sum_i D_i Y_{it}$ and $\bar{Y}_t^0 = N_0^{-1} \sum_i (1 - D_i) Y_{it}$ are the sample means for the treated and untreated groups in period t.

Note that $\hat{\tau}^{DiD}$ is a special case of the class of estimators of the form

$$\hat{\tau}_\beta = (\bar{Y}_{t=1}^1 - \bar{Y}_{t=1}^0) - \beta(\bar{Y}_{t=0}^1 - \bar{Y}_{t=0}^0).$$

The estimator $\hat{\tau}_\beta$ takes the simple difference-in-means between the treated and control groups in period $t = 1$, and then adjusts by a factor β times the difference in means in the pre-treatment period.

We now draw connections between estimators of the form $\hat{\tau}_\beta$ and estimators that apply covariate adjustments in cross-sectional random experiments. Note that under Assumption 2, $Y_{i,t=0} = Y_{i,t=0}(0)$ regardless of i's treatment status. Our setting is thus isomorphic to a cross-sectional randomized experiment in which the outcome of interest is $Y_i = Y_{i,t=1}$ and we have fixed pre-treatment covariates $X_i = Y_{i,t=0}(0)$. In the cross-sectional setting, Lin (2013) and Li and Ding (2017) consider estimators of the form

$$\hat{\tau}(b_0, b_1) = \bar{Y}^1 - \bar{Y}^0 - (\bar{X}^1 - \bar{X})b_1 + (\bar{X}^0 - \bar{X})b_0,$$

where $\bar{Y}^1 = N_1^{-1} \sum_i D_i Y_i$ and the other terms are defined analogously. Observe, however, that the unconditional mean $\bar{X} = N^{-1} \sum_i X_i$ is a weighted average of $\bar{X}^1$ and $\bar{X}^0$, i.e. $\bar{X} = (N_0/N)\bar{X}^0 + (N_1/N)\bar{X}^1$. It then follows from some straightforward algebra that

$$\hat{\tau}(b_0, b_1) = (\bar{Y}^1 - \bar{Y}_0) - \left(\frac{N_0}{N}b_1 + \frac{N_1}{N}b_0\right)(\bar{X}^1 - \bar{X}^0).$$

The estimator $\hat{\tau}(b_0, b_1)$ is thus equivalent to $\hat{\tau}_\beta$ with $\beta = (N_0/N)b_1 + (N_1/N)b_0$.

With this equivalence in hand, it is straightforward to apply the results in Lin (2013) and Li and Ding (2017) to show that i) $\hat{\tau}_\beta$ is unbiased for the ATE for all β, and ii) solve for the efficient coefficient β^* that minimizes the variance of $\hat{\tau}_\beta$.

Proposition 2.1 (Unbiasedness of $\hat{\tau}_\beta$). *Under Assumptions 1 and 2, $\mathbb{E}\left[\hat{\tau}_\beta\right] = \tau_1$ for all β.*

Proof. The proof is immediate from the results in Lin (2013) and Li and Ding (2017) from the analogy to covariate adjustment in randomized experiment, but we provide a short proof for

8

completeness. Observe that $\bar{Y}_{t=1}^1 = N_1^{-1}\sum_i D_i Y_{it=1}(1)$. By Assumption 1, $\mathbb{E}[D_i] = N_1/N$, so

$$\mathbb{E}\left[\bar{Y}_{t=1}^1\right] = \frac{1}{N_1}\sum_i \mathbb{E}[D_i] Y_{it=1}(1) = \frac{1}{N}\sum_i Y_{it=1}(1).$$

By analogous arguments for the other terms in (1), we have that

$$\mathbb{E}\left[\hat{\tau}^{DiD}\right] = \left(\frac{1}{N}\sum_i Y_{it=1}(1) - \frac{1}{N}\sum_i Y_{it=1}(0)\right) - \left(\frac{1}{N}\sum_i Y_{it=0}(1) - \frac{1}{N}\sum_i Y_{it=0}(0)\right)$$

$$= \tau_1 + \left(\frac{1}{N}\sum_i Y_{it=0}(1) - \frac{1}{N}\sum_i Y_{it=0}(0)\right).$$

However, Assumption 2 implies that the second term in the previous display is zero, which gives the desired result. $\square$

Proposition 2.2. *Let β_d be the coefficient on $Y_{it=1}(0)$ from a regression of $Y_{it=1}(d)$ on $Y_{it=0}(0)$ and a constant. Let $\beta^* = (N_0/N)\beta_1 + (N_1/N)\beta_0$. If Assumptions 1 and 2 hold, then*

$$\mathbb{V}ar[\hat{\tau}_{\beta*}] \leq \mathbb{V}ar[\hat{\tau}_\beta] \text{ for all } \beta \in \mathbb{R},$$

with strict inequality for any $\beta \neq \beta^*$ if $\mathbb{V}ar_f[Y_{it=0}] > 0$.

Proof. We have shown that the estimator $\tau_{\beta*}$ is equivalent to the estimator $\tau(\beta_0, \beta_1)$ considered in Lin (2013) and Li and Ding (2017), with $Y_i = Y_{i,t=1}$ and $X_i = Y_{i,t=0}(0)$. It then follows immediately from the results in Lin (2013) and Li and Ding (2017) that $\hat{\tau}_{\beta*}$ has minimal variance. Further, Li and Ding (2017) show that for any (b_0, b_1), $\mathrm{Var}[\hat{\tau}(\beta_0, \beta_1)] < \mathrm{Var}[\hat{\tau}(b_0, b_1)]$ unless $\mathrm{Var}[\hat{\tau}(\beta_0, \beta_1) - \hat{\tau}(b_0, b_1)] = 0$. Thus, $\mathrm{Var}[\hat{\tau}_{\beta*}] < \mathrm{Var}[\hat{\tau}_{\tilde{\beta}}]$ for any $\tilde{\beta}$ unless $\mathrm{Var}[\hat{\tau}_{\beta*} - \hat{\tau}_{\tilde{\beta}}] = 0$. However,

$$\hat{\tau}_{\beta*} - \hat{\tau}_{\tilde{\beta}} = \hat{\tau}(0, \frac{N}{N_0}\beta^*) - \hat{\tau}(0, \frac{N}{N_0}\tilde{\beta}) = \frac{N}{N_0}(\beta^* - \tilde{\beta})\bar{X}^1.$$

Since $\bar{X}^1 = N_1^{-1}\sum_i D_i X_i$ is a simple random sample of size N_1, it has positive variance if $\mathrm{Var}_f[X_i] > 0$. Thus, $\mathrm{Var}[\hat{\tau}_{\beta*}] < \mathrm{Var}[\hat{\tau}_{\tilde{\beta}}]$ if $\mathrm{Var}_f[X_i] > 0$ and $\beta^* \neq \tilde{\beta}$. $\square$

Proposition 2.2 implies that unless the potential outcomes happen to be such that $\frac{N_0}{N}\beta_1 + \frac{N_1}{N}\beta_0 = 1$, the variance of $\hat{\tau}^{DiD}$ is dominated by that of $\hat{\tau}_{\beta*}$.

9

2.3 The plug-in efficient estimator

In practical settings, however, the "oracle" coefficient β^* is not known. Mirroring Lin (2013) in the cross-sectional case, we now show that β^* can be approximated by a plug-in estimate $\hat{\beta}^*$, and the resulting estimator $\hat{\tau}_{\beta*}$ has similar properties to the "oracle" estimator $\hat{\tau}_{\beta}$.

We first describe the construction of the plug-in estimator. Consider a regression of $Y_{it=1}$ on $Y_{it=0}$ and a constant among units with $D_i = 1$. Let $\hat{\beta}_1$ denote the coneffient on $Y_{it=0}$, i.e.

$$\hat{\beta}_1 = \left(\frac{1}{N_1}\sum_i D_i \dot{Y}_{it=0}^2\right)^{-1}\left(\frac{1}{N_1}\sum_i D_i \dot{Y}_{it=0}Y_{it=1}\right),$$

where $\dot{Y}_{it=0} = Y_{it=0} - \frac{1}{N}\sum_i Y_{it=0}$ is the de-meaned pre-treatment outcome. Define $\hat{\beta}_0$ to be the coefficient from the analogous regression among $D_i = 0$ units,

$$\hat{\beta}_0 = \left(\frac{1}{N_0}\sum_i (1-D_i)\dot{Y}_{it=0}^2\right)^{-1}\left(\frac{1}{N_0}\sum_i (1-D_i)\dot{Y}_{it=0}Y_{it=1}\right).$$

Letting $\hat{\beta}^* = (N_0/N)\hat{\beta}_1 + (N_1/N)\hat{\beta}_0$, the estimator $\hat{\tau}_{\hat{\beta}*}$ is then a feasible approximation to $\hat{\tau}_{\beta*}$. It is straightforward to show that the estimator $\hat{\tau}_{\hat{\beta}*}$ is equivalent to the coefficient from the "interacted" regression considered in Lin (2013), i.e. the coefficient on D_i in the ordinary least squares (OLS) regression,

$$Y_{it=1} = \beta_0 + \beta_1 D_i + \beta_2 \dot{Y}_{it=0} + \beta_3 D_i \times \dot{Y}_{it=0} + \epsilon_i. \tag{2}$$

We will now show that when the population is sufficiently large, $\hat{\tau}_{\hat{\beta}*}$ is approximately unbiased for τ and achieves the same variance as the oracle estimator $\hat{\tau}_{\beta*}$. As in Lin (2013) and Li and Ding (2017) and other papers, we consider sequences of populations indexed by m where $N_{1,m}$ and $N_{0,m}$ grow large. For ease of notation, we leave the index m implicit in our notation for the remainder of the paper. We assume the sequence of populations satisfies the following regularity conditions.

Assumption 3 (Sequences of populations). *Let $Y_i(d) = (Y_{it=0}(d), Y_{it=1}(d))'$ for $d = 0, 1$. Let S_0, S_1, and S_{01} denote the finite population variances and covariance of $Y_i(0), Y_i(1)$:*

$$S_d = \frac{1}{N-1}\sum_i (Y_i(d)-\bar{Y}(d))(Y_i(d)-\bar{Y}(d))', \qquad S_{01} = \frac{1}{N-1}\sum_i (Y_i(0)-\bar{Y}(0))(Y_i(1)-\bar{Y}(1))',$$

where $\bar{Y}(d) = N^{-1}\sum_i Y_i(d)$. We assume

(i) $N_1/N \to p_1 \in (0,1)$.

(ii) S_0, S_1, and S_{01} have finite limiting values denoted $S_0^* > 0, S_1^* > 0$, and S_{01}^*.

(iii) $\max_i ||Y_i(d) - \bar{Y}(d)||^2 / N \to 0$ for $d = 0, 1$.

Part (i) of Assumption 3 states that the fraction of treated units converges to a constant strictly between 0 and 1. Part (ii) states that the variance and covariances of the potential outcomes have limits. Part (iii) requires that no single observation dominates the variance of the potential outcomes, and is thus analogous to the familiar Lindeberg condition in sampling contexts. With these assumptions in hand, we can now formally state the sense in which $\hat{\tau}_{\hat{\beta}*}$ is asymptotically unbiased and as efficient as $\hat{\tau}_{\beta*}$.

Proposition 2.3. *Under Assumptions 1, 2, and 3,*

$$\sqrt{N}(\hat{\tau}_{\hat{\beta}*} - \tau) \to_d \mathcal{N}\left(0, \sigma_*^2\right), \tag{3}$$

where $\sigma_*^2 = \lim_{N \to \infty} N \mathbb{V}ar[\hat{\tau}_{\beta*}]$.

Proof. Since $\hat{\tau}_{\beta*}$ is equivalent to the coefficient on D_i from (2), the result follows immediately from Theorem 1 in Lin (2013). Alternatively, this can be viewed as a special case of Proposition 3.2 below. $\square$

To develop intuition for how $\hat{\tau}_{\hat{\beta}*}$ achieves the same asymptotic variance as $\hat{\tau}_{\beta*}$, observe that we can write

$$\begin{aligned}
\hat{\tau}_{\hat{\beta}*} &= \hat{\tau}_0 - \beta^*(\bar{Y}_{t=0}^1 - \bar{Y}_{t=0}^0) + (\hat{\beta}^* - \beta^*)(\bar{Y}_{t=0}^1 - \bar{Y}_{t=0}^0) \\
&= \hat{\tau}_{\beta*} - (\hat{\beta}^* - \beta^*)\left((\bar{Y}_{t=0}^1 - \bar{Y}_{t=0}) - (\bar{Y}_{t=0}^0 - \bar{Y}_{t=0})\right)
\end{aligned}$$

where $\bar{Y}_{t=0} = N^{-1} \sum_i Y_{it=0}(0)$. By standard arguments for finite populations, $\hat{\beta}^* - \beta^*$ and $\bar{Y}_{t=0}^d - \bar{Y}_{t=0}$ are $O_p(N^{-\frac{1}{2}})$. It follows that $\hat{\tau}_{\hat{\beta}*} - \hat{\tau}_{\beta*}$ is the product of $O_p(N^{-\frac{1}{2}})$ terms, and hence is $O_p(N^{-1})$, whereas $\hat{\tau}_{\beta*} - \tau$ is $O_p(N^{-\frac{1}{2}})$. We thus see that the error induced from estimating β^* is of a higher-order than the variation of $\hat{\tau}_{\beta*}$. In sufficiently large populations, the noise induced by estimating β^* is thus negligible. A similar analysis applies to finite-sample bias, which as in Lin (2013) is also $O_p(N^{-1})$.

Remark 1. Building on results in Frison and Pocock (1992), McKenzie (2012) proposes using the coefficient γ_1 from the OLS regression

$$Y_{it=1} = \gamma_0 + \gamma_1 D_i + \gamma_2 \dot{Y}_{it=0} + \epsilon_i, \tag{4}$$

which is sometimes referred to as the Analysis of Covariance (ANCOVA). This differs from the regression representation of the efficient plug-in estimator in (2) in that it omits the

interaction term $D_i \dot{Y}_{it=0}$. Treating $\dot{Y}_{it=0}$ as a fixed pre-treatment covariate, the coefficient $\hat{\gamma}_1$ from (4) is equivalent to the estimator studied in Freedman (2008b,a). The results in Lin (2013) therefore imply that McKenzie (2012)'s estimator will have the same asymptotic efficiency as $\hat{\tau}_{\hat{\beta}*}$ under constant treatment effects. Intuitively, this is because the coefficient on the interaction term in (2) converges in probability to 0. However, the results in Freedman (2008b,a) imply that under heterogenous treatment effects McKenzie (2012)'s estimator may even be less efficient than $\hat{\tau}_0$, which in turns is (weakly) less efficient than $\hat{\tau}_{\hat{\beta}*}$. Relatedly, Wan (2020) proves that $\hat{\beta}_1$ from (2) is asymptotically at least as efficient as $\hat{\gamma}_1$ from (4) in a sampling-based model that assumes normally distributed potential outcomes.

2.4 Variance estimation

To form confidence sets for τ based on (3), one needs to estimate the variance σ_*^2. As is typical in finite population settings, it is not possible to obtain a consistent variance estimate under treatment effect heterogeneity. We show, however, that one can obtain a consistent estimator for an upper bound on the asymptotic variance. The variance estimator is less conservative than the conventional Neyman estimator in that it accounts for heterogeneity that is explained by lagged outcomes.

We begin with the following decomposition of the variance.

Lemma 2.1. *Under Assumptions 1 and 2,*

$$\mathbb{V}ar[\hat{\tau}_{\beta*}] = \frac{1}{N_1}\tilde{S}_1^2 + \frac{1}{N_0}\tilde{S}_0^2 - \frac{1}{N}\tilde{S}_\tau^2,$$

where $\tilde{S}_1^2$ is the finite population variance of $Y_{it=1}(1) - \beta_1 Y_{it=0}(0)$; $\tilde{S}_0^2$ is the finite population variance of $Y_{it=1}(0) - \beta_0 Y_{it=0}(0)$, and $\tilde{S}_\tau^2$ is the finite population variance of $Y_{it=1}(1) - Y_{it=1}(0) - (\beta_1 - \beta_0)Y_{it=0}(0)$.

Proof. Immediate from Example 9.1 in Li and Ding (2017). □

Proposition 2.4. *Let*

$$\tilde{s}_1^2 = \frac{1}{N_1}\sum_i D_i(Y_{it=1} - Y_{it=0}\hat{\beta}_1)^2, \qquad \tilde{s}_0^2 = \frac{1}{N_0}\sum_i (1 - D_i)(Y_{it=1} - Y_{it=0}\hat{\beta}_0)^2.$$

Under Assumptions 3 and 4, $((N/N_1)\tilde{s}_1^2 + (N/N_0)\tilde{s}_0^2) \to_p \sigma_*^2 - \tilde{S}_\tau^{*2}$, where $\tilde{S}_\tau^{*2} = \lim_{N\to\infty} \tilde{S}_\tau^2$.

Proof. Follows as a special case of Lemma 3.7 below. □

Proposition 2.4 shows that the variance estimate $\tilde{s}^2 = (N/N_1)\tilde{s}_1^2 + (N/N_0)\tilde{s}_0^2$ is asymptotically conservative. It is strictly conservative if $\tilde{S}_\tau^{*2} > 0$, meaning that there is positive

asymptotic variance of the "adjusted" treatment effects $\tau_i - (\beta_1 - \beta_0)Y_{it=0}(0)$ – i.e. heterogeneous treatment effects that are not linear functions of lagged outcomes. We note that in a completely randomized experiment, the typical Neyman variance is conservative by the variance of τ_i. Since $\mathbb{V}\mathrm{ar}_f\left[\tau_i - (\beta_1 - \beta_0)Y_{it=0}(0)\right] = \mathbb{V}\mathrm{ar}_f\left[\tau_i\right] - (\beta_1 - \beta_0)^2\mathbb{V}\mathrm{ar}_f\left[Y_{it=0}(0)\right]$, the variance estimator here is less conservative than the usual Neyman variance estimator.

3 Multiple Periods

We now extend the results above to the more complex setting in which there is staggered treatment timing across multiple periods.

3.1 Model

There is again a finite population of N units. We observe data for T periods, $t = 1, .., T$. A unit's treatment status is indexed by $G_i \in \mathcal{G} \subseteq \{1, ..., T, \infty\}$, where G_i corresponds with the first period in which unit i is treated (and $G_i = \infty$ denotes that a unit is never treated). 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 define the vector $Y_i(g) = (Y_{i1}(g), ..., Y_{iT}(g))' \in \mathbb{R}^T$. We let $D_{ig} = 1[G_i = g]$. The observed vector of outcomes for unit i is then $Y_i = \sum_i D_{ig}Y_i(g)$. We treat as fixed the number of units that are first treated at each time g, N_g, and assume that the timing of treatment is random.

Assumption 4 (Random treatment timing). *Let D be the random $N \times |\mathcal{G}|$ matrix with (i, g)th element D_{ig}. Then $\mathbb{P}(D = d) = (\prod_{g \in \mathcal{G}} N_g!)/N!$ if $\sum_i d_{ig} = N_g$ for all g, and zero otherwise.*

Remark 2 (Stratified Treatment Assignment). For simplicity, we consider the case of unconditional random treatment timing. In some settings, the treatment timing may be randomized among units with some shared observable characteristics (e.g. counties within a state). In such cases, the methodology developed below can be applied to form efficient estimators for each stratum, and the stratum-level estimates can then be pooled to form aggregate estimates for the population.

As in the two-period model, we also assume that the treatment has no causal impact on the outcome in periods before it is implemented.

Assumption 5 (No anticipation). *For all i, $Y_{it}(g) = Y_{it}(g')$ for all $g, g' > t$.*

Note that this assumption does not restrict the possible dynamic effects of treatment – that is, we allow for $Y_{it}(g) \neq Y_{it}(g')$ whenever $t \geqslant min(g, g')$. Rather, we only require that, say, a unit's outcome in period 1 does not depend on whether it was ultimately treated in period 2 or period 3.

3.2 Target Parameter

Following Athey and Imbens (2018), we define $\tau_{it,gg'} = Y_{it}(g) - Y_{it}(g')$ to be the causal effect of switching the treatment date from date g' to g on unit i's outcome in period t. We define $\tau_{t,gg'} = N^{-1} \sum_i \tau_{it,gg'}$ to be the average treatment effect (ATE) of switching treatment from g' to g on outcomes at period t. We will consider scalar estimands of the form

$$\theta = \sum_{t,g,g'} a_{t,gg'} \tau_{t,gg'} \tag{5}$$

i.e. weighted sums of the average treatment effects of switching from treatment g' to g. Researchers will often be intersted in weighted averages of the $\tau_{t,gg'}$, in which case the $a_{t,gg'}$ will sum to 1, although our results allow for general $a_{t,gg'}$.[4] The results extend easily to vector-valued θ'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 5) 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 $t < min(g, g')$.

Researchers are often interested in the effect of receiving treatment at a particular time relative to not receiving treatment at all. We will define $ATE(t, g) := \tau_{t,g\infty}$ to be the average treatment effect on the outcome in period t of being first-treated at period g relative to not being treated at all. The $ATE(t, g)$ is a close analog to the cohort average treatment effects on the treated considered in Callaway and Sant'Anna (2020) and Sun and Abraham (2020). The main difference is that those papers do not assume random treatment timing, and thus consider treatment effects on the treated population rather than average treatment effects (in a sampling-based framework).

Our framework incorporates a variety of possible summary measures that aggregate the $ATE(t, g)$ across different cohorts and time periods. The following definitions mirror those proposed in Callaway and Sant'Anna (2020) for the $ATT(t, g)$. We define the simple-weighted ATE to be the simple weighted average of the $ATE(t, g)$, where each $ATE(t, g)$ is

---

[4]This allows the possibility, for instance, that θ represents the difference between long-run and short-run effects, so that some of the $a_{t,gg'}$ are negative.

weighted by the cohort size N_g,

$$\theta^{simple} = \frac{1}{\sum_t \sum_{g:g\leqslant t} N_g} \sum_t \sum_{g:g\leqslant t} ATE(t, g).$$

Likewise, we define the cohort- and time-specific weighted averages as

$$\theta_t = \frac{1}{\sum_{g:g\leqslant t} N_g} \sum_{g:g\leqslant t} ATE(t, g) \text{ and } \theta_g = \frac{1}{T - g + 1} \sum_{t:t\geqslant g} ATE(t, g),$$

and introduce the summary parameters

$$\theta^{calendar} = \frac{1}{T} \sum_t \theta_t \text{ and } \theta^{cohort} = \frac{1}{|\mathcal{G}\backslash\infty|} \sum_{g:g\neq\infty} \theta_g,$$

where $|A|$ denotes the cardinality of a set A. Finally, we introduce "event-study" parameters that aggregate the treatment effects at a given lag l since treatment

$$\theta_l^{ES} = \frac{1}{\sum_{g:g+l\leqslant T} N_g} \sum_{g:g+l\leqslant T} ATE(g + l, g).$$

Note that the instantaneous parameter θ_0^{ES} is analogous to the estimand considered in de Chaisemartin and D'Haultfœuille (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).

3.3   Class of Estimators Considered

We now introduce the class of estimators we will consider. Let $\hat{\bar{Y}}_g = N_g^{-1} \sum_i D_{ig} Y_i$ be the sample mean of the outcome for treatment group g, and let $\hat{\tau}_{t,gg'} = \hat{\bar{Y}}_{g,t} - \hat{\bar{Y}}_{g',t}$ be the sample analog of $\tau_{t,gg'}$. We define

$$\hat{\theta}_0 = \sum_{t,g,g'} a_{t,gg'} \hat{\tau}_{t,gg'}$$

which replaces the population means in the definition of θ with their sample analogues.

We will consider estimators of the form

$$\hat{\theta}_\beta = \hat{\theta}_0 - \hat{X}'\beta \tag{6}$$

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

$$\hat{X}_j = \sum_{(t,g,g'):g,g'>t} a^j_{0,gg'} \hat{\tau}_{t,gg'}.$$

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 $\hat{\tau}_{t,gg'} - \hat{\tau}_{t,g\infty}$ for a different combination of (t, g, g') with $t < g, g'$. Alternatively, $\hat{X}$ could be a scalar that takes a weighted average of such differences. The choice of $\hat{X}$ is analogous to the choice of which variables to control for in a simple randomized experiment. In principle, including more covariates (higher-dimensional $\hat{X}$) will improve asymptotic precision, yet including "too many" covariates may lead to over-fitting, leading to poor performance in practice.[5] For now, we suppose the researcher has chosen a fixed $\hat{X}$, and will consider the optimal choice of β for a given $\hat{X}$. We will return to the choice of $\hat{X}$ in the discussion of our Monte Carlo results in Section 4 below.

Several estimators proposed in the literature can be viewed as special cases of the class of estimators we consider with $\beta = 1$ and appropriately-defined scalar $\hat{X}$.

Example 1 (Callaway and Sant'Anna (2020)). For settings where there is a never-treated group ($\infty \in \mathcal{G}$), Callaway and Sant'Anna (2020) consider the estimator

$$\hat{\tau}^{CS}_{tg} = \hat{\tau}_{t,g\infty} - \hat{\tau}_{g-1,g\infty},$$

i.e. a difference-in-differences that compares outcomes between period t and $g - 1$ for the cohort first treated in period g relative to the never-treated cohort. It is clear that $\hat{\tau}^{CS}_{tg}$ can be viewed as an estimator of $ATE(t, g)$ of the form given in (6), with $\hat{X} = \hat{\tau}_{g-1,g\infty}$ and $\beta = 1$. Likewise, Callaway and Sant'Anna (2020) consider an estimator that aggregates the $\hat{\tau}^{CS}_{tg}$, say $\hat{\tau}^{CS}_w = \sum_{t,g} w_{t,g} \hat{\tau}_{t,g\infty}$, which can be viewed as an estimator of the parameter $\theta_w = \sum_{t,g} w_{t,g} ATE(t, g)$ of the form (6) with $\hat{X} = \sum_{t,g} w_{t,g} \hat{\tau}_{g-1,g\infty}$.[6] Similarly, Callaway and Sant'Anna (2020) consider an estimator that replaces the never-treated group with an average over cohorts not yet treated in period t,

$$\hat{\tau}^{CS2}_{tg} = \frac{1}{\sum_{g'>t} N_{g'}} \sum_{g'>t} N_{g'} \hat{\tau}_{t,gg'} - \frac{1}{\sum_{g'>t} N_{g'}} \sum_{g'>t} N_{g'} \hat{\tau}_{g-1,gg'}, \text{ for } t \geq g.$$

---

[5]In principle the vector $\hat{X}$ could also include pre-treatment differences in means of non-linear transformations of the outcome as well; see Guo and Basse (2020) for related results on non-linear covariate adjustments in randomized experiments.

[6]This could also be viewed as an estimator of the form (6) if $\hat{X}$ were a vector with each element corresponding with $\hat{\tau}_{t,g\infty}$ and the vector β was a vector with elements corresponding with $w_{t,g\infty}$.

It is again apparent that this estimator can be written as an estimator of $ATE(t,g)$ of the form in (6), with $\hat{X}$ now corresponding with a weighted average of $\hat{\tau}_{g-1,gg'}$ and β again equal to 1.

Example 2 (Sun and Abraham (2020)). Sun and Abraham (2020) consider an estimator that is equivalent to that in Callaway and Sant'Anna (2020) in the case where there is a never-treated cohort. When there is no never-treated group, Sun and Abraham (2020) propose using the last cohort to be treated as the control. Formally, they consider the estimator of $ATE(t,g)$ of the form

$$\hat{\tau}_{tg}^{SA} = \hat{\tau}_{t,gg_{max}} - \hat{\tau}_{s,gg_{max}},$$

where $g_{max} = \max \mathcal{G}$ is the last period in which units receive treatment and $s < g$ is some reference period before g (e.g. $g-1$). It is clear that $\hat{\tau}_{tg}^{SA}$ takes the form (6), with $\hat{X} = \hat{\tau}_{s,gg_{max}}$ and $\beta = 1$. Weighted averages of the $\hat{\tau}_{tg}^{SA}$ can likewise be expressed in the form (6), analogous to the Callaway and Sant'Anna (2020) estimators.

Example 3 (de Chaisemartin and D'Haultfœuille (2020)). de Chaisemartin and D'Haultfœuille (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 $\hat{\tau}_{tg}^{CS2}$. In particular, they consider a weighted average of the treatment effects estimates for the first period in which a unit was treated,

$$\hat{\tau}^{dCH} = \frac{1}{\sum_{g:g \leqslant T} N_g} \sum_{g:g \leqslant T} N_g \hat{\tau}_{t,g\infty}^{CS2}.$$

It is thus immediate from the previous examplse that their estimator can also be written in the form (6).

Example 4 (Two-period DiD). It may be instructive to consider how the estimators considered in the two-period model from Section 2 fit into the general framework. That model corresponds with $T = 2$ and $\mathcal{G} = \{2, \infty\}$.[7] $\hat{X}$ is simply the difference in sample means for the treatment groups in $t = 1$, $\hat{\tau}_{1,2\infty}$. The DiD estimator in the two-period model thus corresponds with $\hat{\theta}_1$.

Example 5 (TWFE Models). Athey and Imbens (2018) consider the setting with $\mathcal{G} = \{1,...T, \infty\}$. Let $D_{it} = 1[G_i \leqslant t]$ be an indicator for whether unit i is treated by period t.

---

[7]Note that the potential outcomes $Y_{it}(\infty)$ and $Y_{it}(2)$ used in this section correspond with the more traditional $Y_{it}(0)$ and $Y_{it}(1)$ used in Section 2.

Athey and Imbens (2018, Lemma 5) show that the coefficient on D_{it} from the two-way fixed effects specification

$$Y_{it} = \alpha_i + \lambda_t + D_{it}\theta^{TWFE} + \epsilon_{it} \tag{7}$$

can be decomposed as

$$\hat{\theta}^{TWFE} = \sum_t \sum_{\substack{(g,g'): \\ min(g,g') \leqslant t}} \gamma_{t,gg'} \hat{\tau}_{t,gg'} + \sum_t \sum_{\substack{(g,g'): \\ min(g,g') > t}} \gamma_{t,gg'} \hat{\tau}_{t,gg'} \tag{8}$$

for weights $\gamma_{t,gg'}$ that depend only on N_g and N and thus are non-stochastic in our framework. Thus, $\hat{\theta}^{TWFE}$ can be viewed as an estimator of the form (6) for the parameter $\theta^{TWFE} = \sum_t \sum_{(g,g'):min(g,g') \leqslant t} \gamma_{t,gg'} \tau_{t,gg'}$, with $X = -\sum_t \sum_{(g,g'):min(g,g')>t} \gamma_{t,gg'} \hat{\tau}_{t,gg'}$ and $\beta = 1$.

Notation. Recall that the sample treatment effect estimates $\hat{\tau}_{t,gg'}$ are themselves differences in sample means, $\hat{\tau}_{t,gg'} = \bar{Y}_{t,g} - \bar{Y}_{t,g'}$. It follows that we can write

$$\hat{\theta}_0 = \sum_g A_{\theta,g} \hat{\bar{Y}}_g \text{ and } \hat{X} = \sum_g A_{0,g} \hat{\bar{Y}}_g$$

for appropriately defined matrices $A_{\theta,g}$ and A_0 of dimension $1 \times T$ and $M \times T$, respectively. These definitions will be useful in deriving our theoretical results below.

3.4 Efficient "Oracle" Estimation

We now consider the problem of finding the best estimator $\hat{\theta}_\beta$ of the form introduced in (6).

First, it is straightforward to show that $\hat{\theta}_\beta$ is unbiased for any fixed β.

Lemma 3.1 ($\hat{\theta}_\beta$ unbiased)**.** *Under Assumptions 4 and 5, $\mathbb{E}\left[\hat{\theta}_\beta\right] = \theta$ for any $\beta \in \mathbb{R}^M$.*

Proof. By Assumption 4, $\mathbb{E}[D_{ig}] = (N_g/N)$. Hence,

$$\mathbb{E}\left[\hat{\theta}_0\right] = \mathbb{E}\left[\sum_g A_{\theta,g} \frac{1}{N_g} \sum_i D_{ig} Y_i\right] = \sum_g A_{\theta,g} \frac{1}{N_g} \sum_i \mathbb{E}[D_{ig}] Y_i(g) = \sum_g A_{\theta,g} \frac{1}{N_g} \sum_i \frac{N_g}{N} Y_i(g) = \theta.$$

Likewise,

$$\mathbb{E}\left[\hat{X}\right] = \mathbb{E}\left[\sum_g A_{0,g} \frac{1}{N_g} \sum_i D_{ig} Y_i\right] = \sum_g A_{0,g} \frac{1}{N} \sum_i Y_i(g) = \frac{1}{N} \sum_i \sum_g A_{0,g} Y_i(g) = 0,$$

since $\sum_g A_{0,g} Y_i(g) = 0$ by Assumption 5. The result follows immediately from the previous two displays. $\qquad \square$

We now turn our attention to deriving the variance of $\hat{\theta}_\beta$ and solving for the most efficient β. We first introduce some notation. Let $S_g = (N-1)^{-1} \sum_i (Y_i(g) - \bar{Y}(g))(Y_i(g) - \bar{Y}(g))'$ be the finite population variance of $Y_i(g)$ and $S_{gg'} = (N-1)^{-1} \sum_i (Y_i(g) - \bar{Y}(g))(Y_i(g') - \bar{Y}(g'))'$ be the finite-population covariance. To derive the variance of $\hat{\theta}_\beta$, it will be useful to solve for the joint variance of $(\hat{\theta}_0, \hat{X}')'$.

Lemma 3.2. *Under Assumptions 4 and 5,*

$$\mathbb{V}ar\left[\begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}\right] = \begin{pmatrix} \sum_g N_g^{-1} A_{\theta,g} S_g A'_{\theta,g} - S_\theta, & \sum_g N_g^{-1} A_{\theta,g} S_g A'_{0,g} \\ \sum_g N_g^{-1} A_{0,g} S_g A'_{\theta,g}, & \sum_g N_g^{-1} A_{0,g} S_g A'_{0,g} \end{pmatrix} =: \begin{pmatrix} V_{\hat{\theta}_0} & V_{\hat{\theta}_0,X} \\ V_{X,\hat{\theta}_0} & V_X \end{pmatrix},$$

where $S_\theta = \mathbb{V}ar_f\left[\sum_g A_{\theta,g} Y_i(g)\right]$.

Proof. Let $A_{\tau,g} = \begin{pmatrix} A_{\theta,g} \\ A_{0,g} \end{pmatrix}$. Then we can write

$$\hat{\tau} := \sum_g A_{\tau,g} \hat{\bar{Y}}_g = \begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}.$$

Since Assumption 4 holds, we can appeal to Theorem 3 in Li and Ding (2017), which implies that $\mathbb{V}ar[\hat{\tau}] = \sum_g N_g^{-1} A_{\tau,g} S_g A'_{\tau,g} - S_\tau$, where $S_\tau = \mathbb{V}ar_f\left[\sum_g A_{\tau,g} Y_i(g)\right]$. The result then follows immediately from expanding this variance, as well as the observation that $S_\tau = \begin{pmatrix} S_\theta & 0 \\ 0 & 0 \end{pmatrix}$, which follows from the fact that $\sum_g A_{0,g} Y_i(g) = \mathbb{E}_f\left[\sum_g A_{0,g} Y_i(g)\right] = 0$ for all i by Assumption 5. $\square$

The variance of $\hat{\theta}_\beta$ then follows immediately.

Corollary 3.1. *Under Assumptions 4 and 5, $\mathbb{V}ar\left[\hat{\theta}_\beta\right] = V_{\hat{\theta}_0} + \beta' V_X \beta - 2V_{\hat{\theta}_0,X}\beta$.*

Having solved for $\mathbb{V}ar\left[\hat{\theta}_\beta\right]$, we now derive the β^* that minimizes the variance.

Proposition 3.1. *Suppose Assumptions 4 and 5 hold and that V_X is full-rank. Let $\beta^* = V_X^{-1} V_{X,\hat{\theta}_0}$, for V_X and $V_{X,\hat{\theta}_0}$ as defined in Lemma 3.2. Then $\mathbb{V}ar\left[\hat{\theta}_{\beta*}\right] = V_{\hat{\theta}_0} - V_{\hat{\theta}_0,X} V_X^{-1} V_{X,\hat{\theta}_0} \leqslant \mathbb{V}ar\left[\hat{\theta}_\beta\right]$ for all $\beta \in \mathbb{R}^M$.*

Proof. First, note that

$$\mathbb{V}ar\left[\hat{\theta}_{\beta*}\right] = \mathbb{V}ar\left[\hat{\theta}_0\right] + \beta^{*'} \mathbb{V}ar\left[\hat{X}\right]\beta^* - 2\,\mathrm{Cov}\left[\hat{\theta}_0, \hat{X}\right]\beta^* = V_{\theta_0} + V_{\hat{\theta}_0,X} V_X^{-1} V_{X,\hat{\theta}_0} - 2V_{\hat{\theta}_0,X} V_X^{-1} V_{X,\hat{\theta}_0},$$

which gives the first equality. Next, observe that

$$V_{\hat{\theta}_\beta} = \mathbb{V}\mathrm{ar}\left[\hat{\theta}_0 - \hat{X}'\beta\right] = \mathbb{E}\left[(\hat{\theta}_0 - \theta - \hat{X}'\beta)^2\right],$$

where we use the fact that $\mathbb{E}\left[\hat{\theta}_0\right] = \theta$ from Lemma 3.1 and that $\mathbb{E}\left[\hat{X}\right] = 0$ from the construction of $\hat{X}$ along with Assumption 5. It is then immediate that β^* is optimal if and only if it solves the least-squares problem

$$\min_\beta \mathbb{E}\left[(\hat{\theta}_0 - \theta - \hat{X}'\beta)^2\right].$$

The solution is

$$\beta^* = \mathbb{E}\left[\hat{X}'\hat{X}\right]^{-1}\mathbb{E}\left[\hat{X}'(\hat{\theta}_0 - \theta)\right] = V_X^{-1}V_{X,\hat{\theta}_0},$$

as needed. □

3.5   Properties of the plug-in estimator

As in Section 2, the efficient estimator $\hat{\theta}_{\beta*}$ is not of practical use since the "oracle" coefficient β^* is not known. We now show that in large populations a feasible plug-in estimator $\theta_{\hat{\beta}*}$ has similar properties to the oracle estimator. In particular, let

$$\hat{S}_g = \frac{1}{N_g - 1}\sum_i D_{ig}(Y_i(g) - \hat{\bar{Y}}_g)(Y_i(g) - \hat{\bar{Y}}_g)$$

and let $\hat{V}_{X,\hat{\theta}_0}$ and $\hat{V}_X$ be the analogs to $V_{X,\hat{\theta}_0}$ and V_X that replace S_g with $\hat{S}_g$ in the definitions. We then define $\hat{\beta}^* = \hat{V}_X^{-1}\hat{V}_{X,\hat{\theta}_0}$. We will now show that the feasible plug-in estimator $\hat{\theta}_{\hat{\beta}*}$ is asymptotically unbiased and as efficient as the oracle estimator $\hat{\theta}_{\beta*}$.

We again consider a sequence of finite populations satisfying certain regularity conditions, analogous to the exercise in Section 2.

Assumption 6.   *(i) For all $g \in \mathcal{G}$, $N_g/N \to p_g \in (0,1)$.*

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

(iii) $\max_{i,g} ||Y_i(g) - \bar{Y}(g)||^2/N \to 0$.

Assumption 6 is analogous to Assumption 3 for the 2-period case. Part (i) requires the probability that treatment begins in period $g \in \mathcal{G}$ converges to a constant strictly between 0 and 1. Part (ii) requires 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.

We now provide two lemmas that characterize the asymptotic joint distribution of $(\hat{\theta}_0, \hat{X}')'$, and show that $\hat{S}_g$ is consistent for S_g^* under Assumption 6. Both results are direct consequences of the general asymptotic results in Li and Ding (2017) for multi-valued treatments in randomized experiments.

Lemma 3.3. *Under Assumptions 4, 5, and 6,*

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_0 - \theta \\ \hat{X} \end{pmatrix} \to_d \mathcal{N}(0, V^*),$$

where

$$V^* = \begin{pmatrix} \sum_g p_g^{-1} A_{\theta,g} S_g^* A_{\theta,g}' - S_\theta^* & \sum_g p_g^{-1} A_{\theta,g} S_g^* A_{0,g}' \\ \sum_g p_g^{-1} A_{0,g} S_g^* A_{\theta,g}' & \sum_g p_g^{-1} A_{0,g} S_g^* A_{0,g}' \end{pmatrix} =: \begin{pmatrix} V_{\hat{\theta}_0}^* & V_{\hat{\theta}_0,X}^* \\ V_{X,\hat{\theta}_0}^* & V_X^* \end{pmatrix},$$

and $S_\theta^* = \lim_{N \to \infty} S_\theta$ (where S_θ is defined in Lemma 3.2).

Proof. As in the proof to Lemma 3.2, we can write

$$\hat{\tau} = \sum_g A_{\tau,g} \hat{\bar{Y}}_g = \begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}.$$

The result then follows from Theorem 5 in Li and Ding (2017), combined with the observation noted in the proof to Lemma 3.2 that $S_\tau = \begin{pmatrix} S_\theta & 0 \\ 0 & 0 \end{pmatrix}$ and hence $S_\tau \to \begin{pmatrix} S_\theta^* & 0 \\ 0 & 0 \end{pmatrix}$. $\square$

Lemma 3.4. *Under Assumptions 4, 5, and 6, $\hat{S}_g \to_p S_g^*$ for all g.*

Proof. Follows immediately from Proposition 3 in Li and Ding (2017). $\square$

It is now straightforward to derive the limiting distribution of $\hat{\theta}_{\hat{\beta}*}$.

Proposition 3.2. *Under Assumptions 4, 5, and 6, $\sqrt{N}(\hat{\theta}_{\hat{\beta}*} - \theta) \to_d \mathcal{N}(0, \sigma_*^2)$, where*

$$\sigma_*^2 = V_{\hat{\theta}_0}^* - V_{X,\hat{\theta}_0}^{*\prime} (V_X^*)^{-1} V_{X,\hat{\theta}_0}^* = \lim_{N \to \infty} N \mathbb{V}ar\left[\hat{\theta}_{\beta*}\right].$$

Proof. Recall that $\hat{\beta}* = \hat{V}_X^{-1} \hat{V}_{X,\hat{\theta}_0}$. It is clear that $\hat{\beta}*$ is a continuous function of $\hat{V}_X$ and $\hat{V}_{X,\theta_0}$, and that $\hat{V}_X$ and $\hat{V}_{X,\theta_0}$ are continuous functions of $\hat{S}_g$. From Lemma 3.4 along with the continuous mapping theorem, we obtain that $\hat{\beta}* \to_p (V_X^*)^{-1} V_{X,\hat{\theta}_0}^*$. Lemma 3.3 together

21

with Slutsky's lemma then give that $\sqrt{N}(\hat{\theta}_{\hat{\beta}*} - \theta) \to_d \mathcal{N}\left(0, V^*_{\hat{\theta}_0} - V^{*\prime}_{X,\hat{\theta}_0}(V^*_X)^{-1}V^*_{X,\hat{\theta}_0}\right)$. From Proposition 3.1, it is apparent that the asymptotic variance of $\hat{\theta}_{\hat{\beta}*}$ is equal to the limit of $N\mathbb{V}\mathrm{ar}\left[\hat{\theta}_{\beta*}\right]$, which completes the proof. $\qquad\square$

3.6 Covariance Estimation

To construct confidence intervals using Proposition 3.1, one requires an estimate of σ^2_*. We first introduce a simple Neyman-style variance estimator that is conservative under treatment effect heterogeneity. We then introduce a refinement to this estimator that adjusts for the part of the heterogeneity explained by $\hat{X}$.

From proposition 3.2 as well as the definition of V^*, we have that

$$\sigma^2_* = \left(\sum_g \frac{1}{p_g} A_{\theta,g} S^*_g A'_{\theta,g} - S^*_\theta\right) - \left(\sum_g \frac{1}{p_g} A_{\theta,g} S^*_g A'_{0,g}\right)\left(\sum_g \frac{1}{p_g} A_{0,g} S^*_g A'_{0,g}\right)^{-1}\left(\sum_g \frac{1}{p_g} A_{\theta,g} S^*_g A'_{0,g}\right).$$

Since S^*_g is consistently estimable (Lemma 3.4), a natural conservative (Neyman-style) variance estimator replaces S^*_g with $\hat{S}_g$ and ignores the $-S^*_\theta$ term. That is, we consider

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

Lemma 3.5. *Under Assumptions 4, 5, and 6, $\hat{\sigma}^2_* \to_p \sigma^2_* + S^*_\theta \leqslant \sigma^2_*$.*

Proof. Immediate from Lemma 3.4 combined with the continuous mapping theorem. $\qquad\square$

Intuitively, the Neyman-type variance estimator proposed above is conservative when there is treatment effect heterogeneity.

When the estimand θ does not involve any treatment effects for the cohort treated in period 1, the estimator $\hat{\sigma}^2_*$ 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); Abadie, Athey, Imbens and Wooldridge (2020), with lagged outcomes playing a similar role to the fixed covariates.[8]

---

[8] Aronow, Green and Lee (2014) provide sharp bounds on the variance of the difference-in-means estimator in randomized experiments, although these bounds are difficult to extend to other estimators and settings like those considered here.

Lemma 3.6. *Suppose that $A_{\theta,g} = 0$ for all $g < g_{min}$. If Assumption 5 holds, then*

$$S_\theta = \mathbb{V}ar_f\left[\tilde{\theta}_i\right] + \left(\sum_{g \geqslant g_{min}} \beta_g\right)'(MS_{g_{min}}M')\left(\sum_{g \geqslant g_{min}} \beta_g\right), \tag{9}$$

where M is the matrix that selects the rows of Y_i corresponding with $t < g_{min}$; $\beta_g = (MS_gM')^{-1}MS_gA'_{\theta,g}$ is the coefficient from projecting $A_{\theta,g}Y_i(g)$ on $MY_i(g)$; and $\tilde{\theta}_i = \sum_{g \geqslant g_{min}} A_{\theta,g}Y_i(g) - \sum_{g \geqslant g_{min}}(MY_i(g))'\beta_g$.

Proof. For any g and functions of the potential outcomes $X_i \in \mathbb{R}^K$ and $Z_i \in \mathbb{R}$, let $\dot{X}_i = X_i - \mathbb{E}_f[X_i]$, $\dot{Z}_i = Z_i - \mathbb{E}_f[Z_i]$, and $\beta_{XZ} = \mathbb{V}ar_f[X_i]^{-1}\mathbb{E}_f\left[\dot{X}_i\dot{Z}_i\right]$. We claim that

$$\mathbb{V}ar_f[Z_i - \beta'_{XZ}X_i] = \mathbb{V}ar_f[Z_i] - \beta'_{XZ}\mathbb{V}ar_f[X_i]\beta_{XZ}.$$

Indeed,

$$\mathbb{V}ar_f[Z_i - \beta'_{XZ}X_i] = \frac{1}{N-1}\sum_i\left(\dot{Z}_i - \beta'_{XZ}\dot{X}_i\right)^2$$

$$= \frac{1}{N-1}\sum_i\dot{Z}_i^2 + \beta'_{XZ}\left(\frac{1}{N-1}\sum_i\dot{X}_i\dot{X}_i'\right)\beta_{XZ} - \beta'_{XZ}\frac{2}{N-1}\sum_i\dot{X}_i\dot{Z}_i$$

$$= \mathbb{V}ar_f[Z_i] + \beta'_{XZ}\mathbb{V}ar_f[X_i]\beta_{XZ} - 2\beta'_{XZ}\mathbb{V}ar_f[X_i]\beta_{XZ}$$

$$= \mathbb{V}ar_f[Z_i] - \beta'_{XZ}\mathbb{V}ar_f[X_i]\beta_{XZ}.$$

The result then follows from setting $Z_i = \sum_{g \geqslant g_{min}} A_{\theta,g}Y_i(g)$ and $X_i = MY_i(g_{min})$, and noting that under Assumption 5, $MY_i(g_{min}) = MY_i(g)$ for all $g \geqslant g_{min}$, and hence $\mathbb{V}ar_f[MY_i(g_{min})] = MS_{g_{min}}M' = MS_gM' = \mathbb{V}ar_f[MY_i(g)]$. $\qquad\square$

The second term on the right-hand side of (9) is consistently estimable, and allows us to obtain a tighter variance estimate. In particular, let

$$\hat{\sigma}^2_{**} = \hat{\sigma}^2_* - \left(\sum_{g > g_{min}} \hat{\beta}_g\right)'\left(M\hat{S}_{g_{min}}M'\right)\left(\sum_{g > g_{min}} \hat{\beta}_g\right),$$

where $\hat{\beta}_g = (M\hat{S}_gM)^{-1}A_{\theta,g}\hat{S}_gM'$.[9] Then $\hat{\sigma}^2_{**}$ is consistent for a tighter upper bound on σ^2_*.

Lemma 3.7. *Suppose that $A_{\theta,g} = 0$ for all $g < g_{min}$ and Assumptions 4-6 hold. Then, $\hat{\sigma}_{**} \to_p \sigma^2_* + S^*_{\tilde{\theta}}$, where $S^*_{\tilde{\theta}} = \lim_{N\to\infty}\mathbb{V}ar_f\left[\tilde{\theta}_i\right]$ for $\tilde{\theta}_i$ defined in Lemma 3.6, and $S^*_{\tilde{\theta}} \leqslant S^*_\theta$.*

---

[9]Assumption 5 implies that $MS_gM' = MS_{g_{min}}M'$ for all $g \geqslant g_{min}$. The term $M\hat{S}_{g_{min}}M'$ can thus be replaced by any convex combination of $M\hat{S}_gM'$ for $g \geqslant g_{min}$; this has no effect on the asymptotic results, but may improve finite sample performance.

Proof. Note that $\hat{\beta}_g$ is a continuous function of $\hat{S}_g$. Lemma 3.4 together with the continuous mapping theorem imply that

$$\left(\sum_{g > g_{min}} \hat{\beta}_g \right)' \left(M\hat{S}_{g_{min}}M' \right) \left(\sum_{g > g_{min}} \hat{\beta}_g \right) - \left(\sum_{g > g_{min}} \beta_g \right)' \left(MS_{g_{min}}M' \right) \left(\sum_{g > g_{min}} \beta_g \right) \to_p 0.$$

The result is then immediate from Lemmas 3.5 and 3.6.

$\square$

3.7 Implications for existing estimators

We now discuss the implications of our results for estimators previously proposed in the literature. As discussed in Examples 1-3 above, the estimators of Callaway and Sant'Anna (2020), Sun and Abraham (2020), and de Chaisemartin and D'Haultfœuille (2020) correspond with the estimator $\hat{\theta}_1$ for an appropriately defined $\hat{X}$. 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 random treatment timing. Since the optimal β^* 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 β^* will typically not be known, our results imply that the plug-in estimator $\hat{\theta}_{\hat{\beta}*}$ will have similar properties in large populations, and thus will be more efficient than the previously-proposed estimators in large populations. We note, however, that the estimators in the aforementioned papers are valid for the ATT in settings where only parallel trends holds but there is not random treatment timing, whereas randomization of treatment timing is necessary for the validity of the efficient estimator.[10] We thus view the results on the efficient estimator as complementary to these estimators considered in previous work.

Similarly, in light of Example 5, our results imply that the TWFE estimator will generally not be the most efficient estimator for the TWFE estimand, θ^{TWFE}. Previous work has argued that the estimand θ^{TWFE} may not be the most economically interesting estimand and may be difficult to interpret (e.g. Athey and Imbens (2018); Borusyak and Jaravel (2016); Goodman-Bacon (2018); de Chaisemartin and D'Haultfœuille (2020)). Our results provide a new and complementary critique of the TWFE specification: even if θ^{TWFE} is the target estimand, estimation via (7) will generally be inefficient in large populations under random treatment timing and no anticipation.

---

[10]The estimator of de Chaisemartin and D'Haultfœuille (2020) can also be applied in settings where treatment turns on and off over time.

4   Monte Carlo Results

We present two sets of Monte Carlo results. In Section 4.1, we conduct simulations in a stylized two-period setting like in Section 2 to illustrate how the efficient estimator compares to the classical difference-in-differences and simple difference-in-means (DiM) estimators. Section 4.2 presents a more realistic set of simulations with staggered treatment timing that is calibrated to the data in Wood et al. (2020a) which we use in our application.

4.1   Two-period Simulations.

Specification. We follow the model in Section 2 in which there are two periods ($t = 0, 1$) and some units are treated in period 1. We first generate the potential outcomes as follows. For each unit i in the population, we draw $Y_i(0) = (Y_{it=0}(0), Y_{it=1}(0))'$ from a $\mathcal{N}(0, \Sigma_\rho)$ distribution, where Σ_ρ has 1s on the diagonal and ρ on the off-diagonal. The parameter ρ is the correlation between the untreated potential outcomes in period $t = 0$ and period $t = 1$. We then set $Y_{it=1}(1) = Y_{it=1}(0) + \tau_i$, where $\tau_i = \gamma(Y_{it=1}(0) - \mathbb{E}_f[Y_{it=1}(0)])$. The parameter γ governs the degree of heterogeneity of treatment effects: if $\gamma = 0$, then there is no treatment effect heterogeneity, whereas if γ is positive then individuals with larger untreated outcomes in $t = 1$ have larger treatment effects. We normalize by $\mathbb{E}_f[Y_{it=1}(0)]$ so that the treatment effects are 0 on average. We generate the potential outcomes once, and treat the population as fixed throughout our simulations. Our simulation draws then differ based on the draw of the treatment assignment vector. For simplicity, we set $N_1 = N_0 = N/2$, and in each simulation draw, we randomly select which units are treated in $t = 1$ or not. We conduct 1000 simulations for all combinations of $N_1 \in \{25, 1000\}$, $\rho \in \{0, .5, .99\}$, and $\gamma \in \{0, 0.5\}$.

Results. Table 1 shows the bias, standard deviation, and coverage of 95% confidence intervals based on the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$, difference-in-differences $\hat{\theta}^{DiD} = \hat{\theta}_1$, and simple differences-in-means $\hat{\theta}^{DiM} = \hat{\theta}_0$. Confidence intervals are constructed as $\hat{\theta}_{\hat{\beta}*} \pm 1.96\hat{\sigma}_{**}$ for the efficient estimator, and analogously for the other estimators.[11] For all specifications and estimators, the estimated bias is quite 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 DiD when auto-correlation of $Y(0)$ is high ($\rho = 0.99$)

---

[11]For $\hat{\theta}_\beta$, we use an analog to $\hat{\sigma}_{**}$, except the unrefined estimate $\hat{\sigma}_*$ for the efficient estimator is replaced with the sample analog to the expression for $\mathbb{V}\text{ar}\left[\hat{\theta}_\beta\right]$ given in Corollary 3.1.

and there are 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 DiM when there is no autocorrelation ($\rho = 0$) and there is no treatment effect heterogeneity ($\gamma = 0$), and thus $\beta^* = 0$ and so DiM is optimal in the class we consider. The plug-in efficient estimator is substantially more precise than DiD and DiM in many other specifications: in the worst specification, the standard deviation of DiD is as much as 1.7 times larger than the plug-in efficient estimator, and the standard deviation of the 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 model is optimal.

				Bias			SD			Coverage		
N_1	N_0	ρ	γ	PlugIn	DiD	DiM	PlugIn	DiD	DiM	PlugIn	DiD	DiM
1000	1000	0.99	0.0	0.00	0.00	−0.00	0.01	0.01	0.04	0.95	0.95	0.95
1000	1000	0.99	0.5	0.00	0.00	−0.00	0.01	0.01	0.06	0.95	0.95	0.95
1000	1000	0.50	0.0	0.00	0.00	0.00	0.04	0.04	0.05	0.94	0.95	0.94
1000	1000	0.50	0.5	0.00	0.00	0.00	0.05	0.05	0.06	0.94	0.95	0.95
1000	1000	0.00	0.0	−0.00	0.00	−0.00	0.04	0.07	0.04	0.95	0.94	0.95
1000	1000	0.00	0.5	−0.01	−0.00	−0.01	0.06	0.07	0.06	0.95	0.95	0.95
25	25	0.99	0.0	0.00	0.00	−0.03	0.04	0.04	0.27	0.94	0.94	0.94
25	25	0.99	0.5	0.00	−0.00	−0.04	0.05	0.08	0.34	0.92	0.93	0.93
25	25	0.50	0.0	−0.01	0.02	−0.02	0.24	0.29	0.26	0.94	0.95	0.94
25	25	0.50	0.5	0.01	0.03	−0.01	0.30	0.32	0.33	0.94	0.95	0.94
25	25	0.00	0.0	−0.03	−0.02	−0.03	0.28	0.38	0.27	0.93	0.95	0.93
25	25	0.00	0.5	−0.04	−0.02	−0.04	0.35	0.42	0.34	0.93	0.94	0.94

Table 1: Bias, Standard Deviation, and Coverage for $\hat{\theta}_{\hat{\beta}*}$, $\hat{\theta}^{DiD}$, $\hat{\theta}^{DiM}$ in 2-period simulations

4.2 Simulations Based on Wood et al. (2020a)

To evaluate the performance of our proposed methods in a more realistic setting, we conduct simulations calibrated to our application to Wood et al. (2020a) in Section 5. 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 5 for more background on the application.

Simulation specification. We calibrate our baseline specification as follows. The number of observations and time periods in the data exactly matches the data from Wood et al.

				SD Relative to Plug-In		
N_1	N_0	ρ	γ	PlugIn	DiD	DiM
1000	1000	0.99	0.0	1.00	1.00	7.09
1000	1000	0.99	0.5	1.00	1.71	7.07
1000	1000	0.50	0.0	1.00	1.13	1.15
1000	1000	0.50	0.5	1.00	1.04	1.15
1000	1000	0.00	0.0	1.00	1.45	1.00
1000	1000	0.00	0.5	1.00	1.31	1.00
25	25	0.99	0.0	1.00	0.99	6.58
25	25	0.99	0.5	1.00	1.47	6.31
25	25	0.50	0.0	1.00	1.21	1.10
25	25	0.50	0.5	1.00	1.08	1.10
25	25	0.00	0.0	1.00	1.35	0.98
25	25	0.00	0.5	1.00	1.22	0.98

Table 2: Ratio of standard deviations for $\hat{\theta}^{DiD}$ and $\hat{\theta}^{DiM}$ relative to $\hat{\theta}_{\hat{\beta}*}$ in 2-period simulations

(2020a) used in our application.[12] We set the untreated potential outcomes $Y_{it}(\infty)$ to match the observed outcomes in the data Y_i (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 Appendix Section A.) In each simulation draw s, we randomly draw a vector of treatment dates $G_s = (G_1^s, ..., G_N^s)$ such that the number of units first treated in periods g matches that observed in the data (i.e. $\sum 1[G_i^s = g] = N_g$ for all g). In total, there are 72 months of data on 7785 officers. There are 48 distinct values of g, with the cohort size N_g ranging from 6 to 642. In an alternative specification, we collapse the data to the yearly level, so that there are 6 time periods and 5 cohorts.

For each simulated data-set, we calculate the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$ for four estimands: the simple weighted average ATE (θ^{simple}); the calendar- and cohort-weighted average treatment effects ($\theta^{calendar}$ and θ^{cohort}), and the instantaneous event-study parameter (θ_0^{ES}). (See Section 3.2 for the formal definition of these estimands). In our baseline specification, we use as $\hat{X}$ the scalar weighted combination of pre-treatment differences used by the Callaway and Sant'Anna (2020, CS) estimator for the appropriate estimand (see Example 1). In the appendix, we also present results for an alternative specification in which $\hat{X}$ is a

---

[12]As in our application, we restrict attention to police officers who remained in the police force throughout the sample period.

vector containing $\hat{\tau}_{t,gg'}$ for all pairs $g, g' > t$. For comparison, we also compute the CS estimator for the same estimand, using the not-yet-treated as the control group (since all units are eventually treated). Recall that for θ_0^{ES}, the CS estimator coincides with the estimator proposed in de Chaisemartin and D'Haultfœuille (2020) in our setting, since treatment is an absorbing state. Confidence intervals are calculated as $\hat{\theta}_{\hat{\beta}*} \pm 1.96\hat{\sigma}_{**}$ for the plug-in efficient estimator and analogously for the CS estimator.

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% confidence intervals 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 CS estimator also is approximately unbiased and has excellent coverage for all of the estimands as well.

Table 4 shows that there are large efficiency gains from using the plug-in efficient estimator relative to the CS estimator. The table compares the standard deviation of the plug-in efficient estimator to that of the CS estimator. Remarkably, using the plug-in efficient estimator reduces the standard deviation by a factor of nearly two for the calendar-weighted average, and by a factor between 1.36 and 1.67 for the other estimands. Since standard errors are proportional to the square root of the sample size, these results suggest that using the plug-in efficient estimator is roughly equivalent to multiplying the sample size by a factor of four for the calendar-weighted average.

Extensions. In Appendix A, we present simulations from an alternative specification where the monthly data is collapsed to the yearly level, leading to fewer time periods and fewer (but larger) cohorts. Both the efficient and CS estimators have very good coverage and minimal bias. The efficient estimator again dominates the CS estimator in efficiency, although the gains are smaller (24 to 30% reductions in standard deviation). The smaller efficiency gains in this specification are intuitive: the CS estimator overweights the pre-treatment periods (relative to the efficient estimator) in our setting, but the penalty for doing this is smaller in the collapsed data, where the pre-treatment outcomes are averaged over more months and thus have lower variance.

In the appendix, we also present results from a modification of our baseline DGP with heterogeneous treatment effects. We again find that the plug-in efficient estimator performs well, with qualititative findings similar to those in the baseline specification, although the standard errors are somewhat conservative as expected.

In the appendix, we also conduct simulation results using a modified version of the plug-

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	0.00	0.93	0.27	0.29
PlugIn	cohort	0.00	0.92	0.24	0.24
PlugIn	ES0	0.01	0.94	0.26	0.27
PlugIn	simple	0.00	0.92	0.22	0.22
CS	calendar	0.00	0.94	0.55	0.55
CS	cohort	-0.01	0.95	0.41	0.41
CS/CdH	ES0	0.01	0.94	0.36	0.36
CS	simple	-0.01	0.96	0.41	0.40

Table 3: Results for Simulations Calibrated to Wood et al. (2020a)

Note: This table shows results for the plug-in efficient and Callaway and Sant'Anna (2020) estimators in simulations calibrated to Wood et al. (2020a). The estimands considered are the calendar-, cohort-, and simple-weighted average treatment effects, as well as the instantaneous event-study effect (ES0). The Callaway and Sant'Anna (2020) estimator for ES0 corresponds with the estimator in de Chaisemartin and D'Haultfœuille (2020). Coverage refers to the fraction of the time a nominal 95% confidence interval includes the true parameter. 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.

Estimand	Ratio of SDs
calendar	1.92
cohort	1.67
ES0	1.36
simple	1.82

Table 4: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator

Note: This table shows the ratio of the standard deviation of the Callaway and Sant'Anna (2020) estimator relative to the plug-in efficient estimator, based on the simulation results in Table 3.

in efficient estimator in which $\hat{X}$ is a vector containing all possible comparisons of cohorts g and g' in periods $t < min(g, g')$. We find poor coverage of this estimator in the monthly specification, where the dimension of $\hat{X}$ is large relative to the sample size (1987, compared with $N = 7785$), and thus the normal approximation derived in Proposition 3.2 is poor. By contrast, when the data is collapsed to the yearly level, and thus the dimension of $\hat{X}$ constructed in this way is more modest (10), the coverage for this estimator is good, and it offers modest efficiency gains over the scalar $\hat{X}$ considered in the main text. These findings align with the results in Lei and Ding (2020), who show that covariate-adjustment in cross-sectional experiments yields asymptotically normal estimators when the dimensions of the

covariates is $o(N^{-\frac{1}{2}})$ (and certain regularity conditions are satisfied). We thus recommend using the version of $\hat{X}$ with all potential comparisons only when its dimension is small relative to the square root of the sample size.

Finally, we repeat the same exercise for the other outcomes used in our application (use of force and sustained complaints). We again find that the plug-in efficient estimator has minimal bias, good coverage properties, and is substantially more precise than the CS estimator for nearly all specifications (with reductions in standard deviations by a factor of over 3 for some specifications). The one exception to the good performance of the plug-in efficient estimator is the calendar-weighted average for sustained complaints when using the monthly data: the coverage of CIs based on the plug-in efficient estimator is only 79% in this specification. Two distinguishing features of this specification are that the outcome is very rare (pre-treatment mean 0.004) and the aggregation scheme places the largest weight on the earliest three cohorts, which were small (sizes 17,15,26). This finding aligns with the well-known fact that the central limit theorem may be a poor approximation in finite samples with a binary outcome that is very rare. The plug-in efficient estimator again has good coverage (94%) when considering the annualized data where the cohort sizes are larger. We thus urge some caution in using the plug-in efficient estimator (or any procedure based on a normal approximation) when cohort sizes are small (<30) and the outcome is rare (mean < 0.01); in such settings, we recommend collapsing the data to a higher level of aggregation before using the plug-in estimator.

5 Application to Wood et al. (2020a)

5.1 Background

Reducing police misconduct and use of force is an important policy objective. Wood et al. (2020a) studied the Chicago Police 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.[13] Wood et al. (2020a) found large and statistically significant impacts of the program on complaints and sustained complaints against police officers and on officer use of force. However, in the process of preparing the analysis for this paper, we discovered a statistical error in the original analysis of Wood et al. (2020a): their cohort-level analysis

---

[13]See the Supplement to Wood et al. (2020a) for discussion of some concerns regarding non-compliance, particularly towards the end of the sample. We explore robustness to dropping officers trained in the last year in Appendix Figure 4. The results are qualitatively similar, although with smaller estimated effects on use of force.

failed to normalize for the fact that cohorts trained on different days were of varying sizes (see Wood et al. (2020b) for details). In Wood et al. (2020b), we collaborated with the authors of Wood et al. (2020a) to correct this error, re-analyzing the data using the procedure proposed by Callaway and Sant'Anna (2020). Our re-analysis found no significant effect on complaints or sustained complaints, and borderline significant effects on use of force, although the confidence intervals for all three outcomes included both near-zero and meaningfully large effects.

5.2   Data

We use the same data as in our re-analysis in Wood et al. (2020b), which extends the data used in the original analysis of Wood et al. (2020a) through December 2016. As in Wood et al. (2020b), we restrict attention to the balanced panel of 7,785 who remained in the police force throughout the study period. The data contain the outcome measures (complaints, sustained complaints, and use of force) at a monthly level for the period of 72 months (6 years), with the first cohort trained in month 13 and the final cohort trained in the last month of the sample. The data also contain the date on which each officer was trained.

5.3   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. We estimate the simple-, cohort-, and calendar-weighted average effects described in Section 3.2 and used in our Monte Carlo study. We also estimate the average dynamic effect 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 Callaway and Sant'Anna (2020) estimator as we did in our re-analysis in Wood et al. (2020b). (Recall that for the instantaneous event-study effect, the Callaway and Sant'Anna (2020) and de Chaisemartin and D'Haultfœuille (2020) estimators coincide.)

5.4   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% confidence intervals (CIs) to the mean of the outcome in the 12 months before treatment began. The estimates using the plug-in efficient estimator are substantially more precise than those using the Callaway and Sant'Anna (2020, CS) estimator, with the standard errors ranging from 1.3 to 5.6 times

smaller (see final column of Table 5).

Figure 1: Effect of Procedural Justice Training Using the Plug-In Efficient and Callaway and Sant'Anna (2020) Estimators
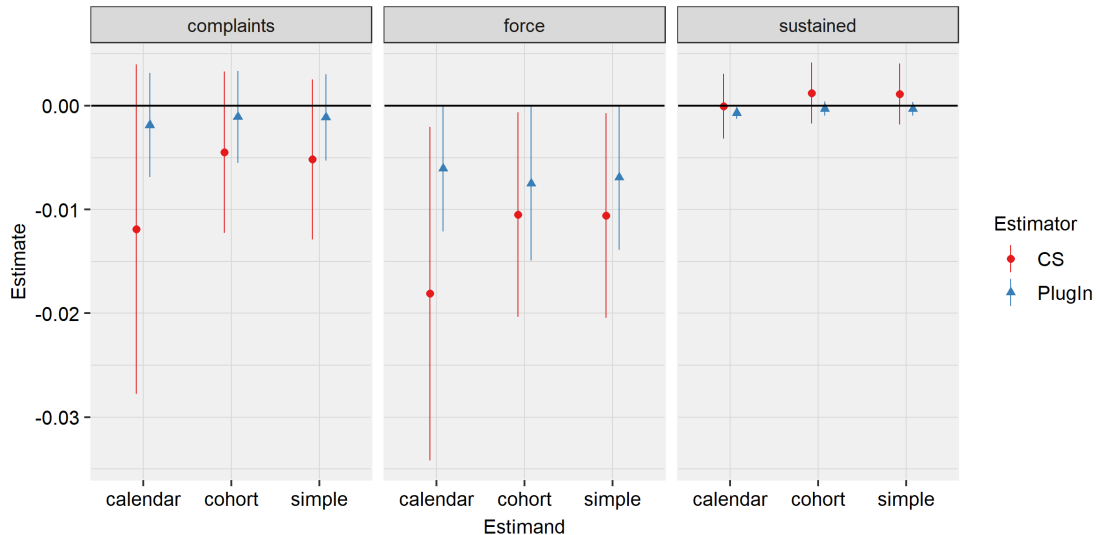
Note: this figure shows point estimates and 95% CIs for the effects of procedural justice training on complaints, force, and sustained complaints using the CS and plug-in efficient estimators. Results are shown for the calendar-, cohort-, and simple-weighted averages.

As in our re-analysis, 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 11%, compared with a bound of 26% using the CS estimator. For use of force, the point estimates are somewhat smaller than when using the CS estimator and the upper bounds of the confidence intervals are all nearly exactly 0. Although precision is substantially higher than when using the CS estimator, the CIs for force still include effects between near-zero and 29% of the pre-treatment 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.[14] If we were to Bonferroni-adjust all of the CIs in Figure 1 for testing nine hypotheses (three outcomes times three aggregations), none of the confidence intervals would rule out zero.

Figure 2 shows event-time estimates for the first two years using the plug-in efficient estimator. (To conserve space, we place the analogous results for the CS estimator in the

---

[14]Recall that the calendar aggregation for sustained complaints was the one specification for which CIs based on the plug-in efficient estimator substantially undercovered (79%), and thus the significant result should be interpreted with some caution.

			Plug-In			CS			
Outcome	Estimand	Pre-treat Mean	Estimate	LB	UB	Estimate	LB	UB	CI Ratio
complaints	simple	0.049	−2%	−11%	6%	−10%	−26%	5%	1.9
complaints	calendar	0.049	−4%	−14%	6%	−24%	−56%	8%	3.2
complaints	cohort	0.049	−2%	−11%	7%	−9%	−25%	7%	1.8
sustained	simple	0.004	−7%	−23%	8%	27%	−44%	97%	4.5
sustained	calendar	0.004	−17%	−30%	−3%	−1%	−75%	73%	5.6
sustained	cohort	0.004	−7%	−22%	9%	29%	−41%	99%	4.4
force	simple	0.048	−15%	−29%	0%	−22%	−43%	−2%	1.4
force	calendar	0.048	−13%	−26%	0%	−38%	−72%	−4%	2.6
force	cohort	0.048	−16%	−31%	−0%	−22%	−43%	−1%	1.3

Table 5: Estimates and 95% CIs as a Percentage of Pre-treatment Means

Note: This table shows the pre-treatment means for the three outcomes. It also displays the estimates and 95% CIs in Figure 1 as percentages of these means. The final columns shows the ratio of the CI length using the CS estimator relative to the plug-in efficient estimator.

appendix.) In dark blue, we present point estimates and pointwise confidence intervals, and in light blue we present simultaneous confidence bands calculated via Bonferroni adjustment. It has been argued that simultaneous confidence bands are more appropriate for event-study analyses since they control size over the full dynamic path of treatment effects (Freyaldenhoven, Hansen and Shapiro, 2019; Callaway and Sant'Anna, 2020). The figure shows that the simultaneous confidence bands include zero for nearly all periods for all three outcomes. Inspecting the results for force more closely, we see that the point estimates are positive (although typically not significant) for most of the first year after treatment, but become consistently negative around the start of the second year from treatment. This suggests that the negative point estimates in the aggregate summary statistics are driven mainly by months after the first-year. Although it is possible that the treatment effects grow over time, this runs counter to the common finding of fadeout in educational programs in general (Bailey, Duncan, Cunha, Foorman and Yeager, 2020) and anti-bias training in particular (Forscher and Devine, 2017).

Finally, in Appendix Figure 4, we present results analogous to those in Figure 1 except removing officers who were treated in the last 12 months of the data. The reason for this

Figure 2: Event-Time Average Effects Using the Plug-In Efficient Estimator

is, as discussed in the supplement to Wood et al. (2020a), there was some non-compliance towards the end of the study period wherein officers who had not already been trained could volunteer to take the training at a particular date. The qualitative patterns after dropping these observations are similar, although the estimates for the effect on use of force are smaller and not statistically significant at conventional levels.

6 Conclusion

This paper considers efficient estimation in settings where the timing of treatment to different units is randomly assigned. Although this random assignment assumption is stronger than the typical parallel trends assumption, it can be ensured by design when the researcher controls the timing of treatment, and is often the justification given for parallel trends in quasi-experimental contexts. We then derive the most efficient estimator in a large class of estimators that nests many existing approaches. Although the "oracle" efficient estimator is not known in practice, we show that a plug-in sample analog has similar properties in large populations, and derive a valid variance estimator for construction of confidence intervals. We find in simulations that the proposed plug-in efficient estimator is approximately unbiased, yields CIs with good coverage, 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.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296.

Aronow, Peter M., Donald P. Green, and Donald K. K. Lee, "Sharp bounds on the variance in randomized experiments," *The Annals of Statistics*, June 2014, *42* (3), 850–871.

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman, and David S. Yeager, "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions:," *Psychological Science in the Public Interest*, October 2020.

Basse, Guillaume, Yi Ding, and Panos Toulis, "Minimax designs for causal effects in temporal experiments with treatment habituation," *arXiv:1908.03531 [stat]*, June 2020. arXiv: 1908.03531.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2016.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, December 2020.

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

Ding, Peng and Fan Li, "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment," *Political Analysis*, 2019, *27* (4), 605–615.

Doleac, Jennifer, "How to Fix Policing," *Neskanen Center*, 2020.

Forscher, Patrick S and Patricia G Devine, "Knowledge-based interventions are more likely to reduce legal disparities than are implicit bias interventions," 2017.

Freedman, David A., "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, 2008, *2* (1), 176–196.

_ , "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Freyaldenhoven, Simon, Christian Hansen, and Jesse Shapiro, "Pre-event Trends in the Panel Event-study Design," *American Economic Review*, 2019, *109* (9), 3307–3338.

Frison, L. and S. J. Pocock, "Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design," *Statistics in Medicine*, September 1992, *11* (13), 1685–1704.

Funatogawa, Takashi, Ikuko Funatogawa, and Yu Shyr, "Analysis of covariance with pre-treatment measurements in randomized trials under the cases that covariances and post-treatment variances differ between groups," *Biometrical Journal*, May 2011, *53* (3), 512–524.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," Working Paper 25018, National Bureau of Economic Research September 2018.

Guo, Kevin and Guillaume Basse, "The Generalized Oaxaca-Blinder Estimator," *arXiv:2004.11615 [math, stat]*, April 2020. arXiv: 2004.11615.

Imai, Kosuke and In Song Kim, "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 2020, (Forthcoming).

Lei, Lihua and Peng Ding, "Regression adjustment in completely randomized experiments with a diverging number of covariates," *Biometrika*, December 2020, (Forthcoming).

Li, Xinran and Peng Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, October 2017, *112* (520), 1759–1769.

Lin, Winston, "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, March 2013, *7* (1), 295–318.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, April 2015, *124*, 1–17.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

McKenzie, David, "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221.

Meer, Jonathan and Jeremy West, "Effects of the Minimum Wage on Employment Dynamics," *Journal of Human Resources*, 2016, *51* (2), 500–522.

Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923, *5* (4), 465–472.

Owens, Emily, David Weisburd, Karen L. Amendola, and Geoffrey P. Alpert, "Can You Build a Better Cop?," *Criminology & Public Policy*, 2018, *17* (1), 41–87.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602 [econ, stat]*, August 2020. arXiv: 2008.00602.

Roth, Jonathan and Pedro H. C. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," *arXiv:2010.04814 [econ, stat]*, January 2021. arXiv: 2010.04814.

Shaikh, Azeem and Panos Toulis, "Randomization Tests in Observational Studies with Staggered Adoption of Treatment," *arXiv:1912.10610 [stat]*, December 2019. arXiv: 1912.10610.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, December 2020.

Słoczyński, Tymon, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, 2020, (Forthcoming).

Wan, Fei, "Analyzing pre-post designs using the analysis of covariance models with and without the interaction term in a heterogeneous study population," *Statistical Methods in Medical Research*, January 2020, *29* (1), 189–204.

Wood, George, Tom R. Tyler, and Andrew V. Papachristos, "Procedural justice training reduces police use of force and complaints against officers," *Proceedings of the National Academy of Sciences*, May 2020, *117* (18), 9815–9821.

_ , _ , _ , Jonathan Roth, and Pedro H.C. Sant'Anna, "Revised Findings for "Procedural justice training reduces police use of force and complaints against officers"," *Working Paper*, 2020.

Xiong, Ruoxuan, Susan Athey, Mohsen Bayati, and Guido Imbens, "Optimal Experimental Design for Staggered Rollouts," *arXiv:1911.03764 [econ, stat]*, November 2019. arXiv: 1911.03764.

A   Additional Simulation Results

This section presents results from extensions to the simulations in Section 4.

Other outcomes.   Tables 6-9 show results analogous to those in the main text, except using the other two outcomes considered in our application (use of force and sustained complaints).

Annualized data.   Tables 10-15 show versions of our simulations results (for all three outcomes) when the data is collapsed to the annual level, so that there are 6 total time periods and 5 cohorts.

Augmented $\hat{X}$.   Table 16 shows results for an alternative version of the efficient estimator where $\hat{X}$ is now a vector that contains the difference in means between cohort g and g' in all periods $t < min(g, g')$. This vector is large relative to sample size in the monthly specification ($dim(\hat{X}) = 1987$, $N = 7785$), which leads to bias and severe undercoverage for the modified plug-in efficient estimator. In the annualized data, the dimension of the modified $\hat{X}$ is modest (10), and the modified efficient estimator has good coverage and yields small efficiency gains (up to 3%) relative to the plug-in efficient estimator considered in the main text.

Heterogeneous Treatment Effects.   Tables 17 and 18 show simulation results for a modification of our baseline specification in which there are heterogeneous treatment effects. In the baseline specification, $Y_i(g) = Y_i(\infty)$ for all g. In the modification, we set $Y_i(g) = Y_i(\infty) + 1[t >= g] \cdot u_i$. The u_i are mean-zero draws drawn from a normal distribution with standard deviation equal to half the standard deviation of the untreated potential outcomes. We draw the u_i once and hold them fixed throughout the simulations, which differ only in the assignment of treatment timing. The results are similar to those for the main specification, although as expected, the standard errors are somewhat conservative (i.e. the mean standard error exceeds the standard deviation of the estimator).

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	0.03	0.94	0.30	0.32
PlugIn	cohort	0.02	0.92	0.28	0.29
PlugIn	ES0	0.01	0.96	0.28	0.28
PlugIn	simple	0.01	0.93	0.26	0.27
CS	calendar	0.03	0.95	0.59	0.60
CS	cohort	0.01	0.96	0.45	0.44
CS/CdH	ES0	0.01	0.96	0.37	0.37
CS	simple	0.01	0.96	0.45	0.44

Table 6: Results for Simulations Calibrated to Wood et al. (2020a) – Use of Force

Note: This table shows results analogous to Table 3, except using Use of Force rather than Complaints as the outcome.

Estimand	Ratio of SDs
calendar	1.88
cohort	1.51
ES0	1.34
simple	1.65

Table 7: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Use of Force

Note: This table shows results analogous to Table 4, except using Use of Force rather than Complaints as the outcome.

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	0.00	0.79	0.06	0.07
PlugIn	cohort	0.00	0.92	0.03	0.03
PlugIn	ES0	0.01	0.95	0.08	0.08
PlugIn	simple	0.00	0.92	0.03	0.03
CS	calendar	0.01	0.95	0.14	0.17
CS	cohort	0.01	0.95	0.11	0.11
CS/CdH	ES0	0.01	0.94	0.11	0.12
CS	simple	0.01	0.96	0.11	0.12

Table 8: Results for Simulations Calibrated to Wood et al. (2020a) – Sustained Complaints

Note: This table shows results analogous to Table 3, except using Sustained Complaints rather than Complaints as the outcome.

Estimand	Ratio of SDs
calendar	2.58
cohort	3.58
ES0	1.42
simple	3.74

Table 9: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Sustained Complaints

Note: This table shows results analogous to Table 4, except using Sustained Complaints rather than Complaints as the outcome.

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	0.11	0.95	1.99	1.96
PlugIn	cohort	0.15	0.95	2.53	2.49
PlugIn	ES0	0.03	0.96	1.65	1.60
PlugIn	simple	0.14	0.95	2.41	2.37
CS	calendar	0.20	0.96	2.65	2.56
CS	cohort	0.26	0.96	3.24	3.13
CS/CdH	ES0	0.04	0.96	2.05	1.98
CS	simple	0.27	0.96	3.17	3.05

Table 10: Results for Simulations Calibrated to Wood et al. (2020a) – Annualized Data

Note: This table shows results analogous to Table 3, except the data is collapsed to the annual level.

Estimand	Ratio of SDs
calendar	1.30
cohort	1.26
ES0	1.24
simple	1.29

Table 11: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Annualized Data

Note: This table shows results analogous to Table 4, except the data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	-0.01	0.94	2.23	2.27
PlugIn	cohort	0.01	0.93	2.81	2.84
PlugIn	ES0	-0.01	0.95	1.76	1.78
PlugIn	simple	0.00	0.93	2.70	2.73
CS	calendar	-0.03	0.94	2.83	2.88
CS	cohort	-0.01	0.94	3.46	3.48
CS/CdH	ES0	-0.05	0.94	2.10	2.11
CS	simple	0.00	0.95	3.41	3.42

Table 12: Results for Simulations Calibrated to Wood et al. (2020a) – Use of Force & Annualized Data

Note: This table shows results analogous to Table 3, except using Use of Force rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

Estimand	Ratio of SDs
calendar	1.27
cohort	1.23
ES0	1.19
simple	1.25

Table 13: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Use of Force & Annualized Data

Note: This table shows results analogous to Table 4, except using Use of Force rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	0.00	0.95	0.43	0.44
PlugIn	cohort	-0.01	0.94	0.53	0.55
PlugIn	ES0	0.01	0.95	0.45	0.45
PlugIn	simple	-0.01	0.94	0.51	0.52
CS	calendar	0.02	0.96	0.69	0.66
CS	cohort	0.03	0.96	0.81	0.78
CS/CdH	ES0	0.01	0.96	0.61	0.60
CS	simple	0.02	0.96	0.80	0.77

Table 14: Results for Simulations Calibrated to Wood et al. (2020a) – Sustained Complaints & Annualized Data

Note: This table shows results analogous to Table 3, except using Sustained Complaints rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

Estimand	Ratio of SDs
calendar	1.51
cohort	1.42
ES0	1.34
simple	1.47

Table 15: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Sustained Complaints & Annualized Data

Note: This table shows results analogous to Table 4, except using Sustained Complaints rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

(a) Monthly Data

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn - Long X	calendar	1.96	0.00	0.13	0.30
PlugIn - Long X	cohort	1.69	0.01	0.04	0.26
PlugIn - Long X	ES0	1.93	0.01	0.21	0.38
PlugIn - Long X	simple	1.78	0.00	0.04	0.23
PlugIn	calendar	0.00	0.93	0.27	0.29
PlugIn	cohort	0.00	0.92	0.24	0.24
PlugIn	ES0	0.01	0.94	0.26	0.27
PlugIn	simple	0.00	0.92	0.22	0.22

(b) Annual Data

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn - Long X	calendar	0.33	0.94	1.93	1.95
PlugIn - Long X	cohort	0.37	0.93	2.47	2.49
PlugIn - Long X	ES0	0.25	0.95	1.59	1.56
PlugIn - Long X	simple	0.38	0.94	2.33	2.36
PlugIn	calendar	0.11	0.95	1.99	1.96
PlugIn	cohort	0.15	0.95	2.53	2.49
PlugIn	ES0	0.03	0.96	1.65	1.60
PlugIn	simple	0.14	0.95	2.41	2.37

Table 16: Performance of Plug-In Efficient Estimator Using Augmented $\hat{X}$

Note: This table shows the bias, coverage, mean standard error, and standard deviation of two versions of the plug-efficient estimator. The estimator with the label "Long X" uses an augmented version of $\hat{X}$ that includes the difference in means between all cohorts g, g' in periods $t < min(g, g')$. The estimator labeled PlugIn uses a scalar $\hat{X}$ such that the CS estimator corresponds with $\beta = 1$, as in the main text. The simulation specification in panel (a) is the baseline specification considered in the main text; in panel (b), the data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	Mean SE	SD
PlugIn	calendar	-0.01	0.95	0.36	0.36
PlugIn	cohort	0.00	0.95	0.27	0.24
PlugIn	ES0	0.01	0.96	0.29	0.27
PlugIn	simple	-0.01	0.96	0.25	0.22
CS	calendar	-0.02	0.95	0.60	0.57
CS	cohort	-0.01	0.96	0.43	0.41
CS/CdH	ES0	0.01	0.95	0.38	0.36
CS	simple	-0.01	0.97	0.43	0.40

Table 17: Results for Simulations Calibrated to Wood et al. (2020a) – Heterogeneous Treatment Effects

Note: This table shows results analogous to Table 3, except the DGP adds heterogeneous treatment effect as described in Section A.

Estimand	Ratio of SDs
calendar	1.88
cohort	1.51
ES0	1.34
simple	1.65

Table 18: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) versus Plug-in Efficient Estimator – Heterogeneous Treatment Effects

Note: This table shows results analogous to Table 4, except the DGP adds heterogeneous treatment effect as described in Section A.
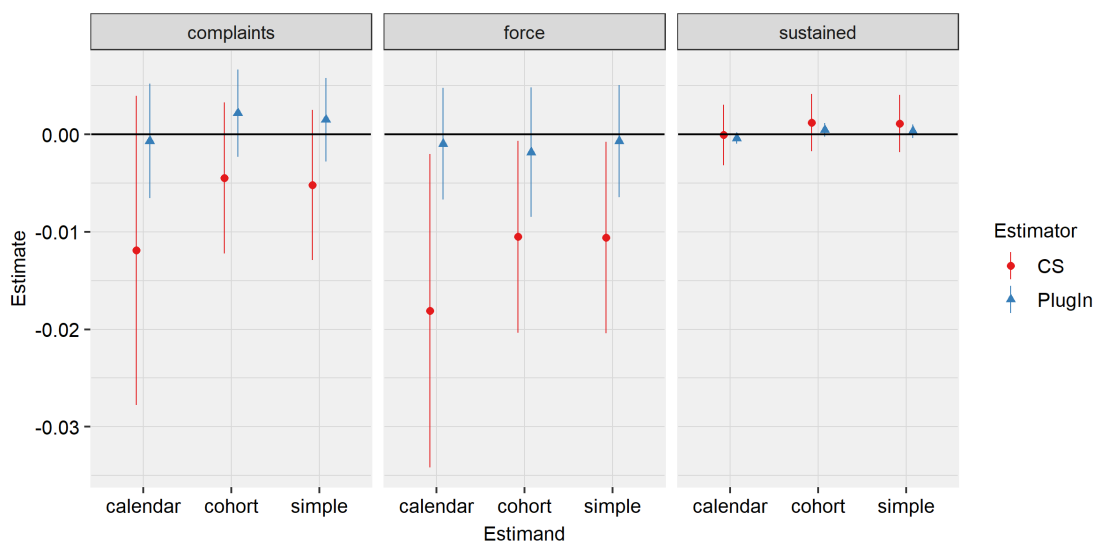
B Additional Tables and Figures

Figure 3: Event-Time Average Effects Using the CS Estimator
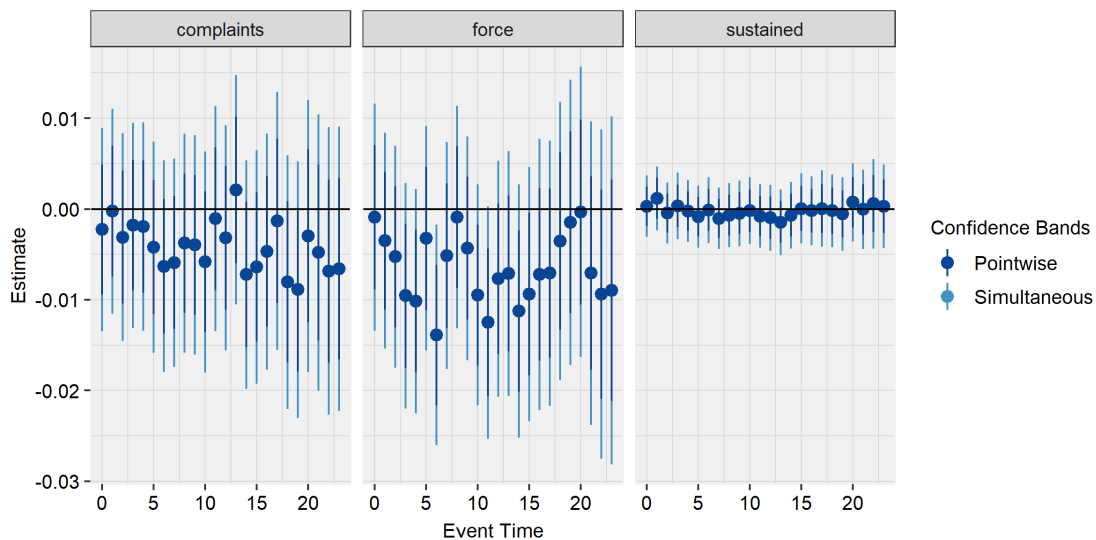
Note: This figure is analogous to Figure 2 except it uses the CS estimator rather than the plug-in efficient.

Figure 4: Effect of Procedural Justice Training Using the Plug-In Efficient and Callaway and Sant'Anna (2020) Estimators – Dropping Late-Trained Officers

Note: This figure is analogous to Figure 1, except we remove from the data officers trained in the last 12 months of the data owing to concerns about treatment non-compliance.